BAYESIAN HIERARCHICAL GAUSSIAN PROCESS MODELS FOR FUNCTIONAL DATA ANALYSIS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Cecilia Ann Earls

August 2014

BAYESIAN HIERARCHICAL GAUSSIAN PROCESS MODELS FOR FUNCTIONAL DATA ANALYSIS

Cecilia Ann Earls, Ph.D.

Cornell University 2014

This dissertation encompasses a breadth of topics in the area of functional data analysis where each function is modeled as a Gaussian process within the framework of a Bayesian hierarchical model. As Gaussian processes cannot be worked with directly in this context, a foundational aspect of this work illustrates that using a finite approximation to each process is sufficient to provide good estimates throughout the entire process. More importantly, it is established that using a finite approximation of a bivariate random process within the estimation procedure also results in providing good estimates throughout the entire bivariate process. With this result, the mean and covariance functions associated with a Gaussian process can be considered as random effects within a Bayesian hierarchical model. Inference for both parameters is based upon their posterior distributions which provide not only estimates of these parameters, but also quantifies variation in these parameters. Here we also propose Bayesian hierarchical models for smoothing, functional linear regression, and functional registration. The registration model introduced here is shown to favorably compare with the best registration methods currently available as measured by the Sobolev Least Squares criterion. Within this registration framework, an Adapted Variational Bayes algorithm is introduced to address the computational costs associated with inference in high-dimensional Bayesian models. With multiple examples, both simulated and using real data, it is shown

that this algorithm results in registered function estimates that closely agree with corresponding estimates obtained from an MCMC sampling scheme. With this algorithm, functional prediction is considered for the first time in a registration context. The final area of inference for functional data that is proposed for the first time here is a combined registration and factor analysis model. This model is shown to outperform currently available registration methods for data in which the registered functions vary in more than one functional direction. The models presented here are applied to several simulated data sets as well as data from the Berkeley Growth Study, functional sea-surface temperature data, and a juggling data set.

BIOGRAPHICAL SKETCH

Cecilia Earls lives in Ithaca, NY with her husband, Chris, and three kids, Ben, Jackson, and Sydney. She received a BS in Mathematics from the University of Minnesota in 1996, and a MS in Statistical Science from Cornell University in 2012.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the opportunities and support afforded to me by Cornell University.

Foremost, I would like to thank my advisor, Dr. Giles Hooker, who early in his career has mastered the art of guiding and supporting PhD students. I am deeply grateful for his time, advice, tolerance, and sense of humor.

I would also like to thank Dr. Martin Wells and Dr. James Booth for serving on my PhD committee. They have both offered some ideas/suggestions that greatly enhanced the scope of my final dissertation. I also extend my thanks to Dr. David Ruppert who first recommended me to Dr. Hooker and whose class in Bayesian analysis greatly influenced my work.

I am grateful to my husband, Chris, for all of the advice and support he has given me throughout this process. Great kudos goes out to my my kids, Ben, Jackson, and Sydney, who are so amazing that they practically raise themselves. I thank all of my family for being patient with me and filling my life with love, purpose, and joy during this crazy time.

My sincerest gratitude goes out to my parents, Dave and Raelene Donarski, who have given me constant support, unconditional love, and have shown me first-hand the joys of both working hard and playing hard. I would like to thank my sister and personal party planner, Felicity Donarski, for helping me take care of countless little details, and also my brother, Eric Donarski, and his family for providing constant sources of entertainment throughout these years.

Also, thank you to all of my friends for their encouragement and support, especially Tami Bennett, Andrea Erven-Victoria, Linda King, Jeff Salipante, Jeanie Cirilano, and Dori Decker.

I am also grateful to all of my friends at CW Taekwondo where I have spent many hours working off stress.

Last, but certainly not least, I thank God who truly makes all things possible.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The primary focus of this dissertation is the flexible use of Gaussian process (GP) distributions in random effect models for functional data analysis. These random effect models are unique in the ease in which they are adapted to a multitude of inferential procedures. All analysis for this dissertation is conducted in a Bayesian environment where inferential procedures are performed for each unknown parameter through the posterior distribution of that parameter. In this context, presented here are novel approaches to statistical inference in the following areas of functional data analysis: non-parametric covariance function estimation, functional smoothing, functional linear regression, functional registration, and combined functional registration and factor analysis. A secondary focus of this dissertation concerns the computational costs of high-dimensional Bayesian hierarchical models. To address these costs, an adapted form of the variational Bayes algorithm is developed to improve computational performance in both the registration and the combined factor analysis and registration models.

Functional data analysis (FDA) is concerned with the analysis of replicated smooth random processes over a continuous domain, most commonly time which we write as $X_1(t), \ldots, X_N(t)$. See Ramsay and Silverman [35] for an overview of models and examples. Many of these methods can be thought of as extensions of multivariate analysis to infinite-dimensional data, combined with smoothing methods to ensure the stability of estimates.

While it is unrealistic to assume that the processes in question are observed exactly at all times, much of the early work in FDA assumed that observations

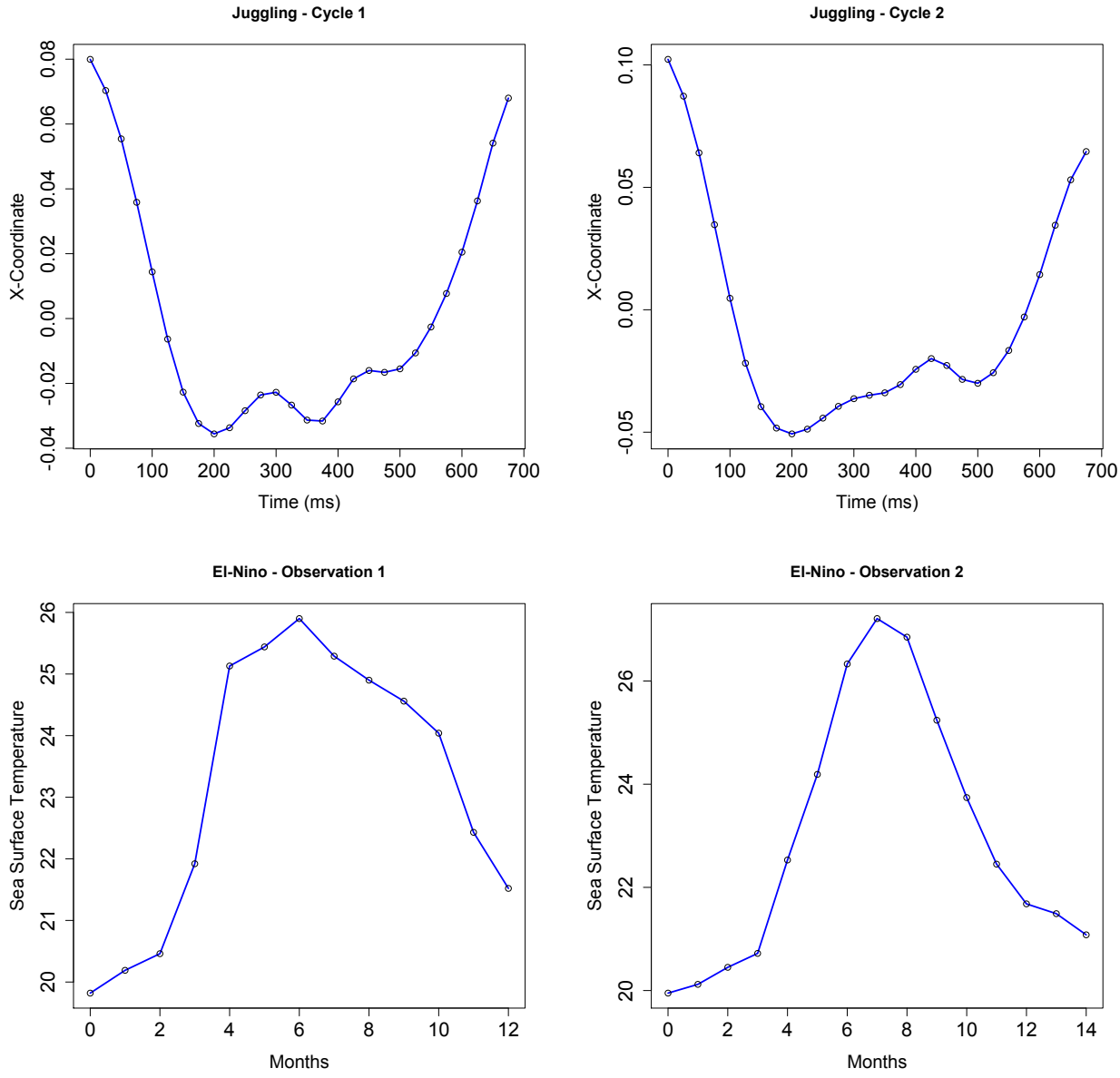

Figure 1.1: Examples of Functional Data. For both examples, the circles are observed data. The lines represent the approximated functional data. **Top Left and Right** The x-coordinate of the right forefinger of Dr. Michael Newton as he juggles, recorded in milliseconds. Each plot represents one juggling cycle. **Lower Right and Left** Each plot contains a "year" of sea-surface temperature recorded monthly.

are precise and frequent enough that pre-smoothing could be employed to obtain representations of the smooth processes. Figure 1.1 contains some examples of functional data used for analysis in this dissertation that fall into this category. The functions in the top row are taken from a data set where the x-coordinate of the right forefinger of a juggler has been recorded over time. Each plot represents one juggling cycle. The bottom row contains two records of sea-surface temperature observed monthly. Each figure here corresponds to one year of sea-surface temperature records, where a year is defined by the time between the two lowest sea-surface temperatures in a 15 month period. A distinguishing feature of both the juggling and sea-surface temperature data sets is that observations are assumed to be taken from the actual process of interest, where each process is observed over a finite subset of the domain. Below we will consider data in which the process of interest cannot be observed directly. The circles in Figure 1.1 correspond to observed data. The solid lines are the "estimated" functional data. In these examples, "estimation" corresponds only to estimating the function between observed time points via linear interpolation of the observations. For more information on these data sets, refer to Sections 4.3 and 3.4.2 respectively.

All functions in this dissertation are characterized by a Gaussian process that provides a non-parametric alternative to the use of a basis/coefficient system to model functions. Specifically, assuming the functions of interest are $X_i(t)$, $i = 1, \ldots, N$, the common assumption for all statistical models presented here is that

$$X_i(t) \mid \mu(t), \Sigma_X(s,t) \quad \sim \quad GP(\mu(t), \Sigma_X(s,t)) \quad s,t \in \mathcal{T} \quad i = 1, \ldots, N \tag{1.1}$$

Similarly, in Section 2.1, an infinite dimensional extension of an inverse-Wishart distribution is introduced to also allow a non-parametric specification of the

covariance function, $\Sigma_X(s,t)$.

A critical aspect of this work is the demonstration that posterior information for our models can be reliably obtained by approximating functional distributions to the evaluation of all parameters at a common set of points followed by linear or bi-linear interpolation between them. This effectively reduces the problem to one of Bayesian estimation in a multivariate latent-vector model. In Gaussian process models with a known covariance function, this approach yields exactly the marginal posterior of the latent processes and their mean at the evaluation points. This is not the case when we must also sample from the posterior of the covariance surface. However, we demonstrate that as the set of evaluation points becomes dense, the difference between adding additional evaluation points and linearly interpolating to obtain posterior values at these evaluation points converges to zero. This serves, first, to demonstrate that our finite-dimensional representation of the posterior is a reliable approximation to the true posterior at the evaluation points and, second, to point to linear or bi-linear interpolation as an efficient means of producing posterior samples at time points not included in the original evaluation points as an alternative to repeating the sampling procedure.

This dissertation extends the current literature on functional data analysis by providing a complete Bayesian framework for inference in FDA that includes non-parametric modeling and inference for functional parameters in a single estimation process. This then allows variance due to the estimation of mean and variance parameters to be incorporated within inferential procedures and provides a framework for inference in more complex models in which latent

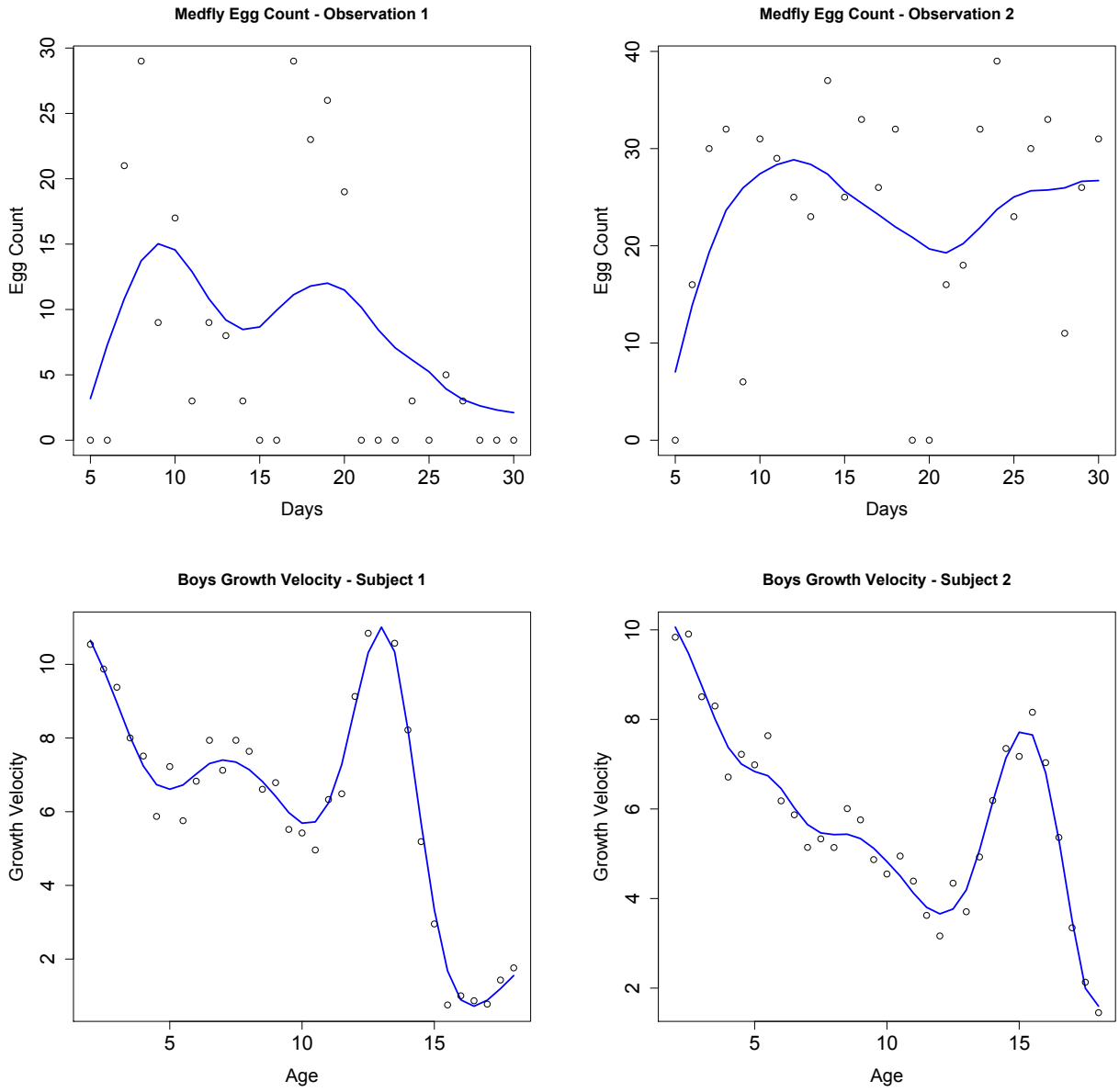

Figure 1.2: Examples of Noisy Functional Data. For both examples, the circles are observed data. **Top Left and Right** Observations in both the left and right figures are the number of eggs laid daily by distinct medflies. The solid lines are the estimated latent functional data representing a smooth biological process that drives the egg laying for each fly. **Lower Left and Right** Each set of observations represent the velocity of growth for a boy from the Berkeley study recorded semiannually. It is assumed that these data are observed noisily (here the noise is simulated). The solid lines are the estimated latent growth velocity functions for each boy.

functional processes need not be directly observed at all.

Recent attention has been given to situations in which each replicate process is observed noisily, infrequently, and possibly at irregularly spaced intervals yielding a model of the form $Y_i(t_{ij}) = X_i(t_{ij}) + \epsilon_{ij}$; see Yao et. al. [51] for pioneering work. This framework essentially makes each postulated process an infinite dimensional latent variable. In this dissertation, data recorded with noise are considered in the context of covariance estimation, functional regression, and functional registration models.

Figure 1.2 contains examples of latent functional data. In the top two figures, observations are daily egg-laying records for two different medflies. The solid line in both illustrations is the estimated latent functional biological process that is assumed to drive egg-laying. The bottom two figures are the growth velocity records for two boys from the Berkeley Growth study in which growth was measured semiannually. For this data set, the data are assumed to be "observed" noisily. However, for the purpose of illustrating our model's ability to estimate the noise process, the noise here is simulated. The actual growth velocity curves are assumed to be smooth and are considered as latent functional data. The estimates of the latent growth velocity processes for these boys are denoted by the solid lines. For both the medfly and Berkeley Growth data sets, the estimated latent processes are determined through the use of a hierarchical Bayesian GP model designed to smooth the observed data. For more information on these data sets see Sections 2.4 and 3.5.3 respectively.

The properties established here for using finite approximations to infinite dimensional distributions in functional data models allow the covariance function, $\Sigma_X(s,t)$ in (1.1) to be considered as a random effect in these models. Thus,

in the Bayesian environment in which inference is performed, not only can the covariance function be estimated, but the posterior distribution of this function can be approximated. In related work, Yao et. al. [51] proposed smoothing a method-of-moments representation of the covariance surface obtained at pairs of observation time points and reconstructing the latent functions via a principal components analysis (PCA) of this surface. This approach was also followed in Goldsmith et. al. [15] in the context of regressing a response on the estimated scores. Crainiceanu and Goldsmith [11] presented a Bayesian version of this regression in which the uncertainty in the latent PCA scores is accounted for, but relied on a pre-estimate of the covariance surface via methods similar to Yao et. al. [51]; none of these methods incorporate uncertainty in the covariance estimate into inferential procedures. Covariance has also been estimated via a spline representation with a penalized log-likelihood, Kauermann and Wegener [18] and also Cai and Yuan [5]. Within Bayesian methods, Linde [25] considers the covariance of a set of spline coefficients to represent functional data, and Kaufman and Sain [19] employs a class of covariance functions that are characterized by a small number of parameters. Nguyen and Gelfand [29] take a Bayesian approach to classifying functions that, similar to our models, are noisily observed; two major areas in which our models are different from Nyugen and Gelfand are 1) Nyugen and Gelfand use canonical components to model the latent functions and 2) they model their covariance functions parametrically which implicitly assumes there are no long-range dependencies in the functions.

Our approach differs from those currently available in both incorporating all parameters in a GP model within a single hierarchical Bayesian analysis and in removing restrictions on the class of covariance surfaces that are used. We demonstrate that smoothness assumptions usually made directly on the $X_i(t)$

can be effectively reproduced within priors on the mean function and covariance surface. We also include priors on smoothing parameters, avoiding the need for cross validation and show that this has the effect of providing additional numerical stability to our Gibbs sampling procedure. Behseta et. al. [1] proposed a hierarchical model to describe variation in the covariance function, but required a separate smoothing procedure from which plug-in estimators are derived.

The development of these methods opens a path to the use of latent functional variables within complex statistical models. In this dissertation, we demonstrate their application to the setting of functional linear regression model

$$z_i = \alpha + \int \beta(t) X_i(t) dt + \epsilon_i \tag{1.2}$$

in which the $X_i(t)$ are observed only noisily and modeled as coming from a latent Gaussian process. We demonstrate that the estimation of the covariance parameter in this process noticeably increases the posterior variance of $\mu(t)$, the common mean of the GP that describes the latent functional covariates, indicating that inferential procedures that do not account for this may have poor coverage probabilities. Another example of a latent Gaussian process model in the context of registration can be found in Section 3.5.

This dissertation also introduces a novel approach to functional data registration within a Bayesian hierarchical model. Registration can be defined as any algorithm that aligns functions in a way that eliminates all phase variability between functions. Without registration, basic summary statistics such as the sample mean and covariance are less interpretable as time variation between significant features in the functions tends to dampen the amplitude variation in

these features. Furthermore, the average timing of significant features may also be of interest and is difficult to obtain under traditional methods of analyzing functional data. For the following discussion, we make the following notational assumptions

$X(t)$:= Unregistered function

$f(t)$:= "target function" used to align functions

$h(t)$:= warping function that defines a map from the unregistered function to the registered function

$X(h(t))$:= $X(t)$ registered by the warping function $h(t)$

$\mathcal{T} = [t_1, t_p]$:= functional domain

A typical example of functional data that contains significant phase variability can be found in Figure 1.3. The left panel contains functional observations of the velocity of growth for 39 boys from the Berkeley Growth study. Significant features in these data include the peak growth rates found on average around ages 7 and 14. However, even a simple estimate of the mean amplitude of these peaks cannot be obtained by analyzing the unregistered data set. After registration, these features are aligned and estimates of the mean realizations of these features can be determined using traditional methods. The estimated registered functions are seen in the right panel of Figure 1.3.

There has been much recent interest in proper ways to define and measure registration as well as in developing registration methods with desirable statistical properties. The evolution of registration dates back to Sakoe and Chiba [39] where the authors use a dynamic programming algorithm for landmark

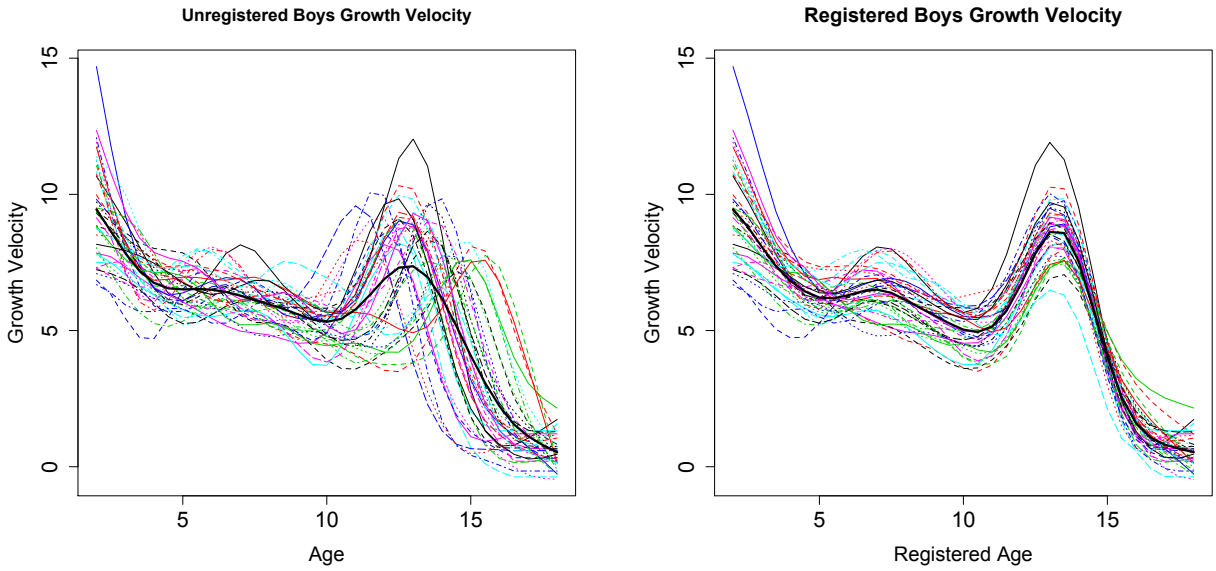

Figure 1.3: Berkeley Data - Boys Growth Velocity. **Top Left** Unregistered boys velocity data functions, $X_i(t)$, $i = 1, \ldots, 39$. **Top Right** Boys velocity functions registered by a Bayesian hierarchical GP model, $X_i(\hat{h}_i(t))$, $i = 1, \ldots, 39$.

registration. Landmark registration is one of the most restricted forms of registration where significant landmarks of each function are identified and aligned. Since 1978, significant advancements in registration procedures either focus on improvements in registration or an effort to combine registration with other inferential procedures.

Landmark registration was again considered by Kneip and Gasser [20] and Kneip and Gasser [21]. Wang and Gasser [48] introduced a new cost function for functional registration. A significant advancement in registration literature can be traced to Silverman [41] and Ramsay and Li [34] where the authors introduce global registration procedures, and Ramsey considers the use of a flexible family of monotone warping functions. Parametric and B-spline base warping functions are considered by Brumback and Lindstrom [4] and Gervini and Gasser

[13], respectively. Nonparametric maximum likelihood approaches to registration are considered by both Ronn [38] and Gervini and Gasser [14]. A moments based approach to registration is introduced by James [17]. Tang and Müller [44] propose pairwise curve synchronization. The first Bayesian approach to registration can be found in Telesca and Inoue [45]. Registration to principal components is considered by Kneip and Ramsay [22]. Finally, with regard to improvements in registration, the recent work by Srivistava, et.al. [43] offers the most comprehensive framework for registration to date.

In Srivistava, et.al. [43], the authors' most significant contribution is in defining a metric for comparing functions that is invariant under warping. The proposed metric is a generalized form of the Fisher-Rao Riemannian metric that has the following property. For two functions, $X_i(t)$ and $X_j(t)$, and any properly defined warping function, $h(t)$, distance defined by Fisher-Rao metric, d_{FR}, is such that

$$d_{FR}(X_i(t), X_j(t)) = d_{FR}(X_i(h(t)), X_j(h(t)))$$

This metric is intractable to work with directly. However, the authors additionally show that if you define the square root velocity transformation (SRVF), $q(t)$, of function, $X(t)$, as

$$q(t) = \frac{X'(t)}{\mid X'(t) \mid^{\frac{1}{2}}}$$

the Fisher-Rao Riemannian metric on the space of functions is equivalent to the L^2 metric on the space of SRVFs. The resulting analysis compares all functions in the SRVF space where the authors' focus is in determining an estimate of the target function, $f(t)$, that has the following property. In the SRVF space, for each function $X_i(t)$ and its SRVF representation $q_i(t)$, $i = 1 \dots N$, $X_i(t)$ is best aligned

using the warping function, $h_i(t)$, where

$$h_i(t) = \underset{h(t)}{\operatorname{argmin}} \| f(t) - \left(q_i(h(t)) \sqrt{h'(t)}\right) \| \tag{1.3}$$

This target function up to a warping, which we will denote as $\tilde{f}(t)$, is determined iteratively in the SRVF space. The final estimate of $f(t)$ is $\tilde{f}(h(t))$ where $h(t)$ is chosen so that the mean of the estimated warping functions determined by the temporary target function, $\tilde{f}(t)$, $\overline{h_\cdot(t)} = \frac{1}{N}\sum_{i=1}^{N} \tilde{h}_i(t)$, is the identity, i.e. $\overline{h_\cdot(t)} = t$ for all $t \in \mathcal{T}$. The last step of their algorithm is a final determination of the warping functions, $h_i(t)$, $i = 1 \ldots N$, using criterion (1.3).

The registration model proposed in this dissertation is framed in a Bayesian environment where a formal metric is unnecessary. However, similar to Srivistava, et.al. [43] we define a target function that is iteratively determined within the warping procedure. This aspect of both models sets them apart from most current registration methods and likely contributes to estimates of similar quality being obtained using these two registration methods. In Section 3.3.1 is a comparison of registration results from that of Srivistava, et.al. [43] and those determined by our proposed Bayesian hierarchical model.

Much of the focus in combining registration with other types of inference in one model has been in the area of functional data clustering and registration. Current work in this area can be found in Liu and Yang [27], Sangalli et. al. [40], and also a Bayesian approach in Zhang and Telescar [52]. Additionally, recent work by Rakêt et. al. [31] includes a model for functional smoothing and registration. While these extensions to registration procedures offer additional tools for functional data analysis, they tend to focus less on high-quality registration.

This dissertation proposes one model that addresses both areas of development in registration procedures. First, a hierarchical Bayesian model is pro-

posed to simultaneously register and smooth functions. It will be demonstrated that this registration model has alignment properties similar to Srivistava, et.al. [43]. Then, this model is extended to encompass functional prediction for future outcomes of a partially recorded (unregistered) function. To our knowledge, this is the first time functional prediction is addressed in a registration framework.

The registration model proposed here is rooted in the work by Ramsay and Li [34]. In this paper, the authors propose a continuous registration method based on a flexible family of non-parametric warping functions. This initial work by Ramsay and Li was then expanded upon in *Functional Data Analysis*, Ramsay and Silverman [36], where the authors propose two different criteria to align functions. The first criterion determines the warping function, $h(t)$, that minimizes the distance between a target function $f(t)$ and the aligned function, $X(h(t))$, with some smoothing requirements on $h(t)$. The second criterion is based on the following construction. Define $\mathbf{f}$ and $\mathbf{X}(\mathbf{h})$ as vectors of the target function and the registered function evaluated over the time points $\mathbf{t} = (t_1, \ldots, t_p)'$. Furthermore let $\mathbf{X} = \begin{bmatrix} \mathbf{f} & \mathbf{X}(\mathbf{h}) \end{bmatrix}$ be the $p \; x \; 2$ matrix of these evaluations. Then, minimizing the smallest eigenvalue of $\mathbf{X}'\mathbf{X}$ with respect to the warping function, $h(t)$, implicitly determines the best warping function by determining $h(t)$ so that these functions vary primarily in one functional direction. In the registration model proposed in Section 3.1, estimates of the warping functions also are determined in part by penalizing variation in the registered functions in directions other than that of the target function. However, in the model proposed here; vertical shifts from the direction of the target function are not penalized, and the target function evolves iteratively within the registration model. Furthermore, in our model, the warping functions are defined to encompass prediction

within our inferential procedures.

In general, warping functions are required to be monotonic increasing. Thus, $h(t)$ must satisfy: for t_j and t_k, $t_j, t_k \in \mathcal{T}$, if $t_k > t_j$, then $h(t_k) > h(t_j)$. Furthermore, here and in most current registration methods, we require: $h(t_1) = t_1$ and $h(t_p) = t_p$. In *Functional Data Analysis*, Ramsay and Silverman [36], Ramsay gives an exact expression for a family of warping functions as follows.

$$h(t) \quad = \quad t_1 + (t_p - t_1)\frac{\int_{t_1}^{t} e^{W(s)}ds}{\int_{t_1}^{t_p} e^{W(s)}ds} \tag{1.4}$$

This definition has the nice property of satisfying the required restrictions on $h(t)$ without any restrictions on the function $W(t)$. The one drawback to this definition of a warping function is that the relationship between $h(t)$ and $W(t)$ is unidentifiable. For instance, let $W(t) = t$, then $W(t)$ and $W(t) + C$, for any constant C are both be mapped to the same function $h(t)$.

Alternatively, in the registration model proposed in this dissertation, $h(t)$ is defined:

$$h(t) \quad = \quad t_1 + \int_{t_1}^{t} e^{w(s)}ds \quad t \in \mathcal{T}$$

where $w(t)$ must satisfy

$$t_p = t_1 + \int_{t_1}^{t_p} e^{w(s)}ds$$

This defines an identifiable relationship between $h(t)$ and $w(t)$ that is necessary for predicting future values of $h(t)$, based on an estimate of $h(t)$ that has been partially determined to a point in the interior of the domain, $\mathcal{T}$. Further details of the prediction model based upon this definition of a warping function can be found in Section 3.4.

In addition to combining data registration and smoothing, the initial registration model presented in this dissertation is extended in Chapter 4 to encompass more flexible registration assumptions. In this context, registration and factor analysis are performed within one hierarchical model. There is no previous work that combines registration and factor analysis; however, in Kneip and Ramsay [22], the authors also consider registration where the aligned functions are assumed to contain variation in more than one functional direction. In their paper, Kneip and Ramsay register functions using an iterative algorithm that updates the PCA decomposition used to register functions in each iteration. This model can be seen as an extension of Ramsay's Procrustes registration method for traditional functional data analysis. In Section 3.3.1, we demonstrate how our initial registration model results in better alignment than Ramsay's Procrustes method.

In addition to covariance function estimation, smoothing, functional linear regression, functional registration, factor analysis, and functional prediction in a registration environment; this dissertation addresses the computational issues associated with high-dimensional Bayesian hierarchical models. To this end, we propose an alternative algorithm to variational Bayes approximation that can be used for models in which the full conditional distributions of a subset of the parameters are not from a known parametric family. We call this the Adapted Variational Bayes (AVB) algorithm and it is applicable to both the registration and the combined factor analysis and registration models presented here.

The general organization of this dissertation is as follows. Chapter 2 presents our foundational work in the use of infinite dimensional distributions for FDA. Inference for the covariance function under the assumption that the functional

data of interest are noisily observed is also found in Section 2.1. The second chapter also presents a functional linear regression model in the context of latent Gaussian processes with a demonstration using medfly data. A Bayesian hierarchical model for registration is introduced in Chapter 3. In this chapter our registration procedure is compared to current methods and applied to simulated data sets, the Berkley Growth data, and sea-surface temperature data. Also, Section 3.2.1 includes a description of the Adapated Variational Bayes algorithm with a discussion on its convergence properties. In Section 3.4, the AVB algorithm is employed for our prediction model. Chapter 4 introduces the combined registration and factor analysis model. Here, again, this method of registration is compared to that of Srivistava, et.al. [43]. Additionally, this chapter includes an analysis of the juggling data using this model. A discussion of the findings of this dissertation and some areas of future work are contained in Chapter 5. Finally, the appendices contain all of the details regarding the MCMC samplers and the AVB algorithm for these models.

CHAPTER 2

BAYESIAN COVARIANCE ESTIMATION AND INFERENCE IN LATENT GAUSSIAN PROCESS MODELS

2.1 Infinite Dimensional Distributions for Functional Estimation and Regression

2.1.1 Gaussian Process Models

Throughout this dissertation, functional data are represented through GP models that are highly flexible in form while retaining descriptive parameters in the mean and covariance functions. The property that the evaluation of a GP model yields a multivariate normal distribution provides a natural reduction of inference for a GP model to a problem in multivariate analysis. We also demonstrate the reverse property; that we can regain a good approximation to the infinite-dimensional process via linear interpolation.

In this chapter, we assume the following data generation model. The data of interest are latent functional data, $X_i(t), i = 1, \ldots, N$, defined on domain, $\mathcal{T}$, modeled by a Gaussian process, for which we only have noisy observations of each function at a given set of time points, t_j, $j = 1, \ldots, p$. Observations, $Y_i(t_j), i = 1, \ldots, N$, $j = 1, \ldots, p$, are independent Gaussian random variables centered at the value of the latent function $X_i(t)$ at time t_j with variance σ^2.

These assumptions result in the following data and latent process models

$$\mathbf{Y} \mid \mathbf{X}, \sigma^2 \sim \prod_{i=1}^{N} N_p(\mathbf{X}_i, \sigma^2 I_p) \tag{2.1}$$

$$X_i(t) \mid \mu(t), \Sigma_X(s,t) \sim GP(\mu(t), \Sigma_X(s,t)) \quad s,t \in \mathcal{T} \quad i = 1, \ldots, N \tag{2.2}$$

where $\mathbf{Y}$ is the matrix such that the observation for function $X_i(t)$ at time point t_j is in the ith row and the jth column, $\mathbf{X}$ is the matrix of the corresponding means for each entry in $\mathbf{Y}$, $\mathbf{X}_i = (X_i(t_1), \ldots, X_i(t_p))'$ is the vector of evaluations of the functions $X_i(t)$ at time points $\mathbf{t} = (t_1, \ldots, t_p)'$, and $\mu(t)$ and $\Sigma_X(s,t)$ are the mean and covariance functions describing the Gaussian process (GP) that characterizes the latent functions, $X_i(t), i = 1, \ldots, N$.

For notational convenience it is assumed that observations are recorded at time points that are common to all latent processes $X_i(t)$. This is not strictly necessary and observations at irregular time points can be readily accommodated by including the evaluation of functions at unobserved time points as further latent variables; the resulting estimate of $X_i(t)$ at times other than those recorded is obtained through the information provided by the estimated mean and covariance functions of the latent processes. This posterior inference is thus an alternative to the use of PCA as proposed in Yao et. al. [51]. In Section 2.4, we demonstrate the application of these methods when blocks of data are missing.

This chapter focuses on estimating parameters in a GP model for functional data. A particularly powerful aspect of these models and methods is their extension to more complex models that include latent functional data, within Bayesian hierarchical models. For example, in Section 2.3, we add responses from the functional linear regression model (1.2) to our framework in which the coefficient function $\beta(t)$ must also be estimated along with $\mu(t)$, $\Sigma_X(s,t)$, and the latent functional processes $X_i(t)$. This represents the most direct use of latent

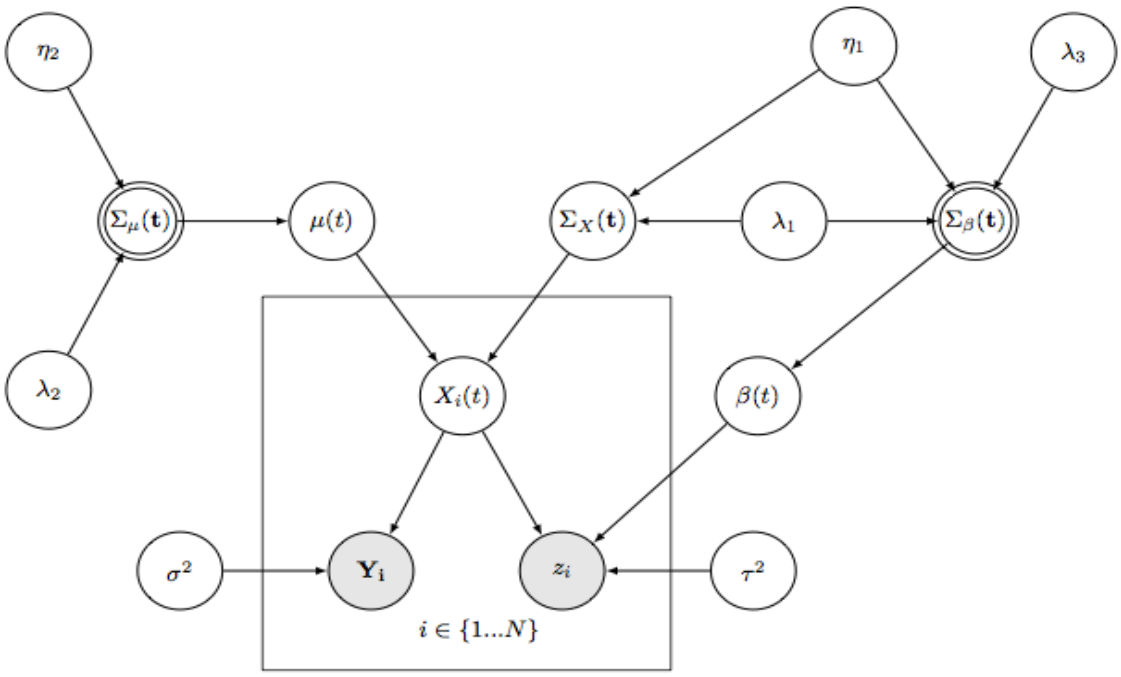

Figure 2.1: Graphical representation of the functional regression model fully specified in Appendix B.2. Shaded circles are observed quantities. Covariance functions defined parametrically as a function of their smoothing parameters are denoted by concentric circles. Specifically, $\Sigma_\mu(\mathbf{t}) = \eta_2^{-1}P_1(\mathbf{t}) + \lambda_2^{-1}P_2(\mathbf{t})$ and $\Sigma_\beta(\mathbf{t}) = (\eta_1\lambda_3)^{-1}P_1(\mathbf{t}) + (\lambda_1\lambda_3)^{-1}P_2(\mathbf{t})$. Here we define $\mathbf{t} = (s,t)'$. Definitions of the bivariate functions P_1 and P_2 can be found in Section 2.2.

GPs within a regression model. In Section 3.5, latent Gaussian process models are considered in the framework of functional registration. Figure 2.1 provides a graphical representation of the functional linear regression model that highlights the dependencies between the observed data and the unknown parameters.

For the latent process and regression models, prior distributions are defined on all unknown parameters. In general, we use prior distributions that are uninformative with the exception of the prior distributions that regularize the pro-

cess. However, even these retain flexibility through the use of diffuse priors on the smoothing parameters that control the regularization process. Section 2.2.2 provides a more thorough discussion of the effect of incorporating prior distributions for the smoothing parameters in this model.

All inference in these models is performed through the posterior sample of each parameter that results from running a Gibbs sampler. Appendix B provides all distributional assumptions as well as the resulting full-conditional distributions necessary to run a Gibbs sampler for these models.

2.1.2 Functional Inference for Parameters Characterized by Infinite Dimensional Distributions

In this section, we provide a definition of our GP model and priors. These models describe infinite dimensional random quantities so that given noisy observations, $\mathbf{Y}_i = (Y_i(t_1) \dots Y_i(t_p))'$, of the latent process, $X_i(t)$, $i = 1, \dots, N$, the distributional assumptions for the observations and latent functions are as given in (2.1) and (2.2).

To this model we append prior distributions for the functional parameters $\mu(t)$ and $\Sigma_X(s,t)$. The prior for $\mu(t)$ is modeled as another Gaussian Process. For $\Sigma_X(s,t)$, we utilize an infinite dimensional extension of an inverse Wishart distribution initially defined by Dawid [8]. This definition uses a nonstandard, but consistent, parametrization which we follow here. The distribution depends on a parameter δ defined as $\delta = \nu - p + 1$, where ν is the degrees of freedom associated with an inverse Wishart distribution defined on a p by p sub-matrix of

the infinite dimensional distribution. Dawid's use of the parameter, δ, allows this parameter to be fixed for any choice of p; in contrast to ν, the degrees of freedom, which is dependent on the dimensionality of the sub-matrix.

In the following, we extend Dawid's definition by allowing an arbitrary scale function $S(s,t)$:

DEFINITION *A bivariate function,* $\Sigma(s,t), s,t \in \mathcal{T}$*, has a functional inverse-Wishart* $(S(s,t),\delta)$ *(FIW) distribution, for* $\delta > 0$*, if the evaluation of* $\Sigma(s,t)$ *over any* $\mathbf{t} \times \mathbf{t}$ *grid has an inverse Wishart* $(\mathbf{S},\delta)$ *distribution with* $p \times p$ *scale matrix,* $\mathbf{S}$*, corresponding to the scale function,* $S(s,t)$*, evaluated over the same grid.*

See Dawid [8] for a complete derivation. This definition of a FIW distribution provides the conditions necessary such that as the dimension, p, of the observations increases, covariance function estimates converge to values of a proper covariance function.

With this definition, we define prior distributions on the parameters of the data and process distributions, (2.1) and (2.2):

$$\begin{aligned} \mu(t) &\sim GP(0, \Sigma_\mu(s,t)) \quad s,t \in \mathcal{T} \\ \Sigma_X(s,t) &\sim FIW(P_X(s,t),\delta) \quad s,t \in \mathcal{T} \\ \sigma^2 &\sim IG(a,b) \end{aligned}$$

where the hyperparameters $P_X(s,t) = \eta_1^{-1}P_1(s,t) + \lambda_1^{-1}P_2(s,t)$ and $\Sigma_\mu(s,t) = \eta_2^{-1}P_1(s,t) + \lambda_2^{-1}P_2(s,t)$ are constructed to provide smoothing information for the mean and latent functions; these are specified in Section 2.2.

In order to obtain a posterior distribution from these definitions, we consider the evaluation of all the $X_i(t)$ and $\mu(t)$ at a common set of time points

$\mathbf{t} = \{t_1, \ldots, t_p\}$ which we will denote repectively as $\mathbf{X}_i$ and $\boldsymbol{\mu}$. In the case of $\Sigma_X(s,t)$, we evaluate on the grid of pairs of time points from $\mathbf{t}$: $[\boldsymbol{\Sigma}_X]_{j,k} = \Sigma_X(t_j, t_k)$. Under the framework above, $\mathbf{X}_i$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}_X$ are described by well-known multivariate distributions for which we can use a Gibbs sampler to obtain posterior distributions. Our methods require the $X_i(t)$ to be evaluated at a common set of time points which must include all the observation time points. However, they need not all be observed at the same time points; the values of the $X_i(t)$ when they are unobserved can still be imputed as additional parameters in our model.

We note that, unlike inference for a single X_i given data $\mathbf{Y}_i = Y_{i1}, \ldots, Y_{ip}$ and parameters μ and Σ_X, inference for μ and Σ_X themselves cannot be immediately undertaken in a point-wise fashion. In particular, the marginal posterior distribution of $\mu(t)$ will depend on the choice of $\mathbf{t}$; similar statements can be made about posterior inference for $\Sigma_X(s,t)$. It is therefore important to show that as $\mathbf{t}$ becomes dense in the time domain, a sensible limit is achieved. We demonstrate this by showing that the difference between including new evaluation points into $\mathbf{t}$ and linearly (or bi-linearly) interpolating estimates using values from the original time points tends to zero as the spacing between observed time points decreases. We note that this also points to potential numerical gains: once an estimate is made for a fine grid $\mathbf{t}$, we can proceed to find estimates for other points via linear interpolation rather than re-running expensive sampling schemes.

In order to carry this out, we make the following assumptions. Note that for this illustration only, the subscript p denotes the dimension of the observations and the usual subscript, "x", for the covariance parameter of the latent functions is suppressed:

A1) The functional parameters are evaluated on a set of equally spaced time

points, $\mathbf{t} = \{t_1, \ldots, t_p\} \subset \mathcal{T}$. We assume this for simplicity; however, it is only necessary that the maximum distance between time points is strictly decreasing.

A2) A functional inverse Wishart prior is defined on the covariance function, $\Sigma(s,t)$, such that $\Sigma(s,t) \sim FIW(P(s,t),\delta)$ and the scale function, $P(s,t)$, is of class C^3 (all third-order partial derivatives are continuous) on $\mathcal{T}$ x $\mathcal{T}$.

A3) $\mathbf{\Sigma}_p$ is the p x p covariance matrix for which the elements consist of the covariance function, $\Sigma(s,t)$, $s,t \in \mathcal{T}$, evaluated over the grid $\mathbf{t}$ x $\mathbf{t}$.

A4) The p-dimensional vector, $\mathbf{f}_p$, is a finite approximation for a functional parameter (e.g. a latent function or the functional mean), $f(t), t \in \mathcal{T}$, where $\mathbf{f}_p = (f(t_1), \ldots, f(t_p))'$.

A5) Conjugate priors are defined on $\mathbf{\Sigma}_p$ and $\mathbf{f}_p$ to employ a Gibbs sampler where the resulting full-conditionals corresponding to parameters, $\mathbf{\Sigma}_p$ and $\mathbf{f}_p$ are

$$\begin{aligned} \mathbf{\Sigma}_p &\sim IW(\mathbf{S}_p, \delta) \\ \mathbf{f}_p &\sim N_p(\boldsymbol{\mu}_p, \mathbf{C}_p) \end{aligned}$$

where $\mathbf{S}_p, k, \boldsymbol{\mu}_p$, and $\mathbf{C}_p$ are known parameters determined by the observations, priors, and current iteration of the sampler.

In addition, we make the following definitions.

D1) $\mathbf{t}_u \subset \mathcal{T}$ is an arbitrary set of r unobserved time points.

D2) $\mathbf{S}_{p+r}$ is the scale matrix corresponding to the scale function evaluated over the grid of observed and unobserved time points, $\{\mathbf{t}, \mathbf{t}_u\}$ x $\{\mathbf{t}, \mathbf{t}_u\}$, such that an associated draw of the covariance function over this grid, $\mathbf{\Sigma}_{p+r}$, is from the

distribution, $\mathbf{\Sigma}_{p+r} \sim IW(\mathbf{S}_{p+r}, \delta)$. Define $\mathbf{S}_{p+r,l}$ as the bi-linear approximation to the scale matrix, $\mathbf{S}_{p+r}$, defined by bi-linear interpolation from the values of $\mathbf{S}_{p+r}$ associated with the observed time points $\mathbf{t}$. Furthermore, if $\mathbf{S}_{p+r} = \mathbf{U}'\mathbf{U}$, define $\mathbf{U}_l$ as the approximation to $\mathbf{U}$ where all entries as a function of the elements of $\mathbf{S}_{p+r}$ are now a function of the corresponding entries in $\mathbf{S}_{p+r,l}$.

D3) $\boldsymbol{\mu}_{p+r,l}$ and $\mathbf{C}_{p+r,l}$ are linear and bi-linear approximations to $\boldsymbol{\mu}_{p+r}$ and $\mathbf{C}_{p+r}$ respectively such that $\mathbf{f}_{p+r} \sim N_{p+r}(\boldsymbol{\mu}_{p+r}, \mathbf{C}_{p+r})$.

PROPOSITION 1 *Suppose the assumptions, A1-A5, hold. A draw from the distribution of* $\mathbf{\Sigma}_{p+r,l} = \mathbf{U}_l'\mathbf{A}_{p+r}\mathbf{U}_l$ *is such that*

$$\| \mathbf{\Sigma}_{p+r,l} - \mathbf{\Sigma}_{p+r} \|_2 \xrightarrow{p} 0$$

where $\mathbf{A}_{p+r} \sim IW(\mathbf{I}_{p+r}, \delta)$.

PROOF. The random matrix $\mathbf{\Sigma}_{p+r} \sim IW(\mathbf{S}_{p+r}, \delta)$ can be represented as

$$\mathbf{\Sigma}_{p+r} = \mathbf{U}'\mathbf{A}_{p+r}\mathbf{U}$$

where $\mathbf{U}$ is the upper triangular matrix of the Cholesky decomposition of the scale matrix, $\mathbf{S}_{p+r}$, and $\mathbf{A}_{p+r} \sim IW(\mathbf{I}_{p+r}, \delta)$. Thus, we can simulate draws from the distribution of $\mathbf{\Sigma}_{p+r}$ in the following way. First draw $\mathbf{A}_{p+r}$ from an $IW(\mathbf{I}_{p+r}, \delta)$ distribution and use this draw to construct a draw from $\mathbf{\Sigma}_{p+r} = \mathbf{U}'\mathbf{A}_{p+r}\mathbf{U}$. We further define $\mathbf{\Sigma}_{p+r,l} = \mathbf{U}_l'\mathbf{A}_{p+r}\mathbf{U}_l$ as an approximation to this draw.

We will rely on the following results to show $\mathbf{\Sigma}_{p+r,l} \xrightarrow{p} \mathbf{\Sigma}_{p+r}$ w.r.t. the L^2 norm.

R1) The matrix $\mathbf{U}_l$, as a function of the linearly approximated scale matrix, $\mathbf{S}_{p+r,l}$, is such that $\| \mathbf{U}_l - \mathbf{U} \|_2 = O(\frac{1}{p})$.

R2) $\| \mathbf{A}_{p+r} \|_2 = \lambda_{max}$, where λ_{max} is the largest eigenvalue of $\mathbf{A}_{p+r}$.

R3) $\lim_{p\to\infty} P(\lambda_{max} \leq \frac{p}{c}) = 1$, where c is some positive fixed constant. This bound is derived from the following convergence property of the smallest eigenvalue of a high-dimensional Wishart matrix by Silverstein [42].

Let $\mathbf{A}_{p+r}^{-1} \sim W(I_{p+r}, \nu)$ define a Wishart distribution with degrees of freedom, ν. Define λ_{min} as the smallest eigenvalue of $\mathbf{A}_{p+r}^{-1}$. Under the condition that $\lim_{\nu\to\infty} \frac{p+r}{\nu} = \gamma \in (0, 1]$, $\frac{1}{\nu}\lambda_{min} \xrightarrow{a.s.} (1 - \gamma^{\frac{1}{2}})^2$ (Silverstein, 1985).

Note, Silverstein's condition is satisfied under the definition of a FIW distribution. In particular, if $A(s,t) \sim FIW(I(s,t), \delta)$. By definition, the marginal distribution of any $(p + r) \times (p + r)$ submatrix, $\mathbf{A}_{p+r}$, of $A(s,t)$ evaluated over the grid $\{\mathbf{t}, \mathbf{t}_u\}$ x $\{\mathbf{t}, \mathbf{t}_u\}$ is $\mathbf{A}_{p+r} \sim IW(\mathbf{I}_{p+r}, \delta = \nu - p - r - 1)$. Thus, $\mathbf{A}_{p+r}^{-1} \sim W(\mathbf{I}_{p+r}, \delta = \nu - p - r - 1)$ and under the additional requirement that $\delta > 0$, $\lim_{\nu\to\infty} \frac{p+r}{\nu} \in (0, 1]$.

We now have for any $\epsilon > 0$,

$$
\begin{aligned}
\lim_{p\to\infty} P(\| \mathbf{\Sigma}_{p+r,l} - \mathbf{\Sigma}_{p+r} \|_2 > \epsilon) &= \lim_{p\to\infty} P(\| \mathbf{U}_l' \mathbf{A}_{p+r} \mathbf{U}_l - \mathbf{U}' \mathbf{A}_{p+r} \mathbf{U} \|_2 > \epsilon) \\
&= \lim_{p\to\infty} P(\| (\mathbf{U}_l - \mathbf{U})' \mathbf{A}_{p+r} (\mathbf{U}_l - \mathbf{U}) \|_2 > \epsilon) \\
&\leq \lim_{p\to\infty} P(\| \mathbf{U}_l' - \mathbf{U}' \|_2 \| \mathbf{A}_{p+r} \|_2 \| \mathbf{U}_l - \mathbf{U} \|_2 > \epsilon) \\
&= \lim_{p\to\infty} P(\lambda_{max} > O(p^2)) \\
&= 0 \qquad \square
\end{aligned}
$$

This convergence property also holds when draws from the distribution of $\mathbf{\Sigma}_p$ are bilinearly approximated to estimate values in $\mathbf{\Sigma}_{p+r}$ corresponding to the unobserved time points, $\mathbf{t}_u$. Let $\mathbf{\Sigma}_{p+r}^l$ represent a bilinearly approximated

draw from the distribution of $\mathbf{\Sigma}_{p+r}$. Then if $\mathbf{L}$ is the operator that augments $\mathbf{\Sigma}_p$ with r linearly interpolated columns that correspond to the unobserved time points, $\mathbf{\Sigma}^l_{p+r} = \mathbf{L}'\mathbf{U}'_p\mathbf{A}_p\mathbf{U}_p\mathbf{L}$, where the known p x p scale matrix $\mathbf{S}_p = \mathbf{U}'_p\mathbf{U}_p$, and $\mathbf{A}_p \sim IW(\mathbf{I}_p, \delta)$. Furthermore, there exists a projection operator, $\mathbf{Q}$, such that $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_p$ and $\mathbf{Q}\mathbf{U}_p\mathbf{L} = \mathbf{U}_l$, where $\mathbf{U}_l$ is defined as above. Replacing $\mathbf{U}_l$ by $\mathbf{Q}\mathbf{U}_p\mathbf{L}$ in the expression for $\mathbf{\Sigma}_{p+r,l}$, $\mathbf{\Sigma}_{p+r,l} = \mathbf{L}'\mathbf{U}'_p\mathbf{Q}'\mathbf{A}_{p+r}\mathbf{Q}\mathbf{U}_p\mathbf{L}$. Recognizing that $\mathbf{Q}'\mathbf{A}_{p+r}\mathbf{Q} \sim IW(\mathbf{I}_p, \delta)$, we conclude that $\mathbf{\Sigma}^l_{p+r} \overset{d}{\sim} \mathbf{\Sigma}_{p+r,l}$.

PROPOSITION 2 *Suppose the assumptions, A1-A5, hold. Then a draw from* $\mathbf{f}_{p+r,l} \sim N_{p+r}(\boldsymbol{\mu}_{p+r,l}, \mathbf{C}_{p+r,l})$ *is such that*

$$\mathbf{f}_{p+r,l} \xrightarrow{d} \mathbf{f}_{p+r}$$

PROOF. The parameters,$\boldsymbol{\mu}_{p+r,l}$ and $\mathbf{C}_{p+r,l}$, have been defined such that

$$\begin{aligned} \lim_{p\to\infty} \boldsymbol{\mu}_{p+r,l} &= \boldsymbol{\mu}_{p+r} \\ \lim_{p\to\infty} \mathbf{C}_{p+r,l} &= \mathbf{C}_{p+r} \end{aligned}$$

It follows that

$$\mathbf{f}_{p+r,l} \xrightarrow{d} \mathbf{f}_{p+r} \qquad \square$$

It is easy to show this convergence property also holds when draws from the distribution of $\mathbf{f}_p$ are linearly interpolated to provide estimates over the time points $\mathbf{t}_u$. As a consequence of these results, the specific choice of $\mathbf{t}$ does not affect the limit as $p \to \infty$. Furthermore, the assumption of smooth hyper-parameters assures good approximation even for relatively small p.

2.2 Parameter Selection for Inverse-Wishart Priors

2.2.1 Scale Functions for Inverse-Wishart Priors

In this section, we describe our specific choice of hyper-parameters for the functional inverse-Wishart distribution. In GP models, smoothness of the $X_i(t)$ is generally guaranteed by the choice of $\Sigma_X(s,t)$ often taking the form of a kernel function $K_h(s-t)$. In the context of functional data analysis, however, $\Sigma_X(s,t)$ must be estimated. Thus, we instead incorporate smoothing information into the scale function, $P_X(s,t)$, for the inverse-Wishart prior and demonstrate that this information is then passed on to posterior distributions for the latent processes $X_i(t)$, $i = 1,\ldots,N$. Below we examine in detail how this scale function is constructed to provide appropriate smoothing information for our models through its finite-dimensional approximation, $\mathbf{P}_X$. In these models, the finite-dimensional approximation to the covariance function has prior distribution, $\mathbf{\Sigma}_X \sim IW(\mathbf{P}_X, \delta)$, where the degrees of freedom are chosen to reflect minimal information.

The following derivation of a penalty on function curvature, provides the basis for the particular form of the scale matrix, $\mathbf{P}_X$, utilized in our models. The derivations below are based on a more general discussion of functional penalties found in Ramsay and Silverman [35].

For function, $X_i(t), t \in \mathcal{T}$, define B as the constraint operator such that $BX_i = [X_i(0), X_i'(0)]'$ and let L be the linear operator such that $LX_i(t) = X_i''(t)$ and $kerL \cap kerB = \emptyset$. Then,
$\| X_i \|^2 = \eta(BX_i)'(BX_i) + \lambda \int (LX_i)^2(t)dt$ defines a penalty on X_i such that larger values

of λ reduce curvature in the latent functions.

Let $\mathbf{L}$ and $\mathbf{B}$ be matrix representations of the operators L and B that define a penalty on the finite approximations to the functions X_i, $i = 1, \ldots, N$, such that

$$\eta(BX_i)'(BX_i) + \lambda \int (LX_i)^2(t)dt \approx \eta \mathbf{X_i'B'BX_i} + \lambda \mathbf{X_i'L'LX_i}$$

In our model, we impose this penalty by defining $\mathbf{P}_X = (\eta \mathbf{B'B} + \lambda \mathbf{L'L})^{-1} = (\eta \mathbf{P}_1^- + \lambda \mathbf{P}_2^-)^{-1}$ as the scale matrix of the inverse Wishart distribution defined for $\mathbf{\Sigma}_X$. We can show that $\mathbf{P}_X$, under this definition, is an approximation to a kernel function characterizing a Hilbert space of real-valued functions, $K(s,t)$, $s, t \in \mathcal{T}$, evaluated over the grid, $\mathbf{t} \times \mathbf{t}$, $\mathbf{t} = \{t_1, \ldots, t_p\}$. If we define $P_1(t_j, t_k)$ as the reproducing kernel for $kerL$ and $P_2(t_j, t_k)$ as the reproducing kernel for $kerB$ evaluated at t_j, $t_k \in \mathbf{t}$,

$$\begin{aligned} \mathbf{P}_X[j,k] &= (\eta \mathbf{B'B} + \lambda \mathbf{L'L})^{-1}[j,k] \\ &= \eta^{-1}(\mathbf{B'B})^-[j,k] + \lambda^{-1}(\mathbf{L'L})^-[j,k] \\ &= \eta^{-1}\mathbf{P}_1[j,k] + \lambda^{-1}\mathbf{P}_2[j,k] \\ &\approx \eta^{-1}P_1(t_j,t_k) + \lambda^{-1}P_2(t_j,t_k) \\ &= K(t_j,t_k) \end{aligned} \tag{2.3}$$

Furthermore, assuming $s, t \in \mathcal{T}$, it can be shown that the reproducing kernel Hilbert space with reproducing kernel, $K(s,t)$, has a dual relationship with a Hilbert space spanned by zero-mean Gaussian random variables, $Z(t)$, $t \in \mathcal{T}$, such that $K(s,t)$ is equivalent to $E(Z(s)Z(t))$ (Wahba [47]). Therefore, the scale matrix, $\mathbf{P}_X$, is also appropriately a covariance function defined on this Hilbert space evaluated over a finite grid of time points.

Now that the scale matrix, $\mathbf{P}_X$, is established as a smoothing agent for the

latent functions that is grounded in a functional environment, we examine the properties of uncertainty quantification in these models. The posterior sample of the covariance matrix, $\mathbf{\Sigma}_X$, provides for a simple way to quantify the uncertainty of the covariance estimate which is otherwise an arduous task, particularly in high dimensions. In Crainiceanu and Goldsmith [11], the underlying covariance function is first estimated using a method of moments approach, smoothed to reflect the functional nature of the data, and then is assumed "known" in the subsequent Bayesian model. Their subsequent approach to Bayesian functional data analysis is based on imputing principal component scores which allows for tractable high-dimensional models. A drawback of this approach is that it does not account for uncertainty in the initial covariance function estimate. In Section 2.4, we demonstrate that under-representing variability in the covariance function estimate can cause an understatement of parameter variability throughout the entire model. Here we suggest a fully Bayesian approach in estimating the covariance function that provides for characterizing this uncertainty.

2.2.2 Automatic Smoothing Parameter Selection

A particular advantage of the Bayesian framework employed in this paper is that smoothing parameters – η and λ above – can be treated as hyper-parameters and included within the same estimation framework as the remaining elements of the model. This is in contrast to approaches such as cross-validation (often followed by subjective adjustment) which requires re-estimation of the model for each value of the smoothing parameters. When more than one smoothing parameter is present, cross-validation results in a difficult multivariate opti-

mization problem; see Ramsay and Silverman [35] and Wood [50] for further discussion and examples.

The selection of smoothing parameters is particularly challenging in our model due to the need to invert a large matrix which can become poorly conditioned when $N < p$. Here we show that the inclusion of smoothing parameters within the Gibbs sampler helps to maintain its numerical stability, where fixing smoothing parameters can lead to problems. This represents a natural and transparent means of ensuring stability as an alternative to methods proposed in Wood [50]. In the models presented in this paper, we assume uninformative Gamma or inverse Gamma priors for all smoothing parameters.

When N, the number of observations is smaller than p, the dimension of the observations, the model must provide smooth estimates of the latent functions to account for the observations providing insufficient information to fully describe function variability between time points. In this situation, the model relies on λ_1 and the bivariate function $P_2(t_j, t_k)$, that are both embedded in the scale function of the prior defined on the covariance function of the latent processes, to provide additional stability. Numerically, the impact of increasing λ_1 can be seen in the following expression for the expected draw of $\mathbf{\Sigma}_X^{-1}$ in the $(m+1)^{st}$ iteration of the sampler.

$$
\begin{aligned}
E(\mathbf{\Sigma}_X^{-1(m+1)}) &= (p+1+N)(\eta_1^{-1(m)}\mathbf{P}_1 + \lambda_1^{-1(m)}\mathbf{P}_2 + \sum_{i=1}^{N}(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})')^{-1} \quad (2.4)\\
&\propto \left(\eta_1^{-1(m)}\sum_{j=1}^{2}\boldsymbol{\nu}_j\boldsymbol{\nu}_j' + \lambda_1^{-1(m)}\sum_{j=3}^{p}\kappa_j^{-1}\boldsymbol{\nu}_j\boldsymbol{\nu}_j' + \sum_{j=1}^{p}(\sum_{i=1}^{N}c_{ij}^{2(m)})\boldsymbol{\nu}_j\boldsymbol{\nu}_j'\right)^{-1} \quad (2.5)\\
&= \sum_{j=1}^{2}\frac{\eta_1^{(m)}}{1+\eta_1^{(m)}\sum_{i=1}^{N}c_{ij}^{2(m)}}\boldsymbol{\nu}_j\boldsymbol{\nu}_j' + \sum_{j=3}^{p}\frac{\lambda_1^{(m)}\kappa_j}{1+\lambda_1^{(m)}\kappa_j\sum_{i=1}^{N}c_{ij}^{2(m)}}\boldsymbol{\nu}_j\boldsymbol{\nu}_j'
\end{aligned}
$$

In (2.4), $\{\boldsymbol{\nu}_j : j = 1 \ldots p\}$ are the eigenvectors of the inverse scale matrix, $\eta_1\mathbf{P}_1^- + \lambda_1\mathbf{P}_2^-$, where $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ represent linear and constant variation while

$\{\boldsymbol{\nu}_j : j = 3, \ldots, p\}$ represent curvature and $\{\eta_1, \eta_1, \{\lambda_1\kappa_j : j = 3, \ldots, p\}\}$ are the corresponding eigenvalues of the inverse scale matrix. Furthermore, for each j, $\sum_{i=1}^{N} c_{ij}^{2(m)}$ represents the variation present in the latent functions in the jth direction in iteration m. This expectation will be numerically unstable when the coefficients of the outer products of the eigenvectors in (2.5) are small. As the coefficients associated with linear and constant variation will remain stable, we are primarily concerned with the behavior of the coefficients associated with directions of curvature. In particular, if $N < p$, we expect the variation in some of the directions associated with curvature to be close to zero. We can see in (2.5), if $\sum_{i=1}^{N} c_{ij}^{2(m)} \approx 0$ for some $j \in \{3, \ldots, p\}$, the coefficient for this j will be reduced to $\lambda_1^{(m)}\kappa_j$ which will move further from zero as $\lambda_1^{(m)}$ increases. Thus, to provide model stability, $\lambda_1^{(m)}$ must be large enough to assure this expected value is non-singular and numerically stable. Looking at an approximate expectation of $\lambda_1^{(m+1)}$, we can see how draws of this parameter reflect the necessary smoothing required to stabilize this model. Assuming a diffuse prior for λ_1 such that the parameters of the associated inverse Gamma distribution are close to zero (and thus can be ignored),

$$E(\lambda_1^{(m+1)}) = \frac{tr(\mathbf{P}_2\boldsymbol{\Sigma}_X^{-1(m+1)})}{(p+1)(p-2)-2} \tag{2.6}$$

We will approximate this expectation by replacing $\boldsymbol{\Sigma}_X^{-1(m+1)}$ in (2.5) by its expected value so that

$$E(\lambda_1^{(m+1)}) \approx \lambda_1^{(m)}\Big(\frac{p+1+N}{p+1}\Big)\frac{1}{p-2}\sum_{j=3}^{p}\frac{1}{1+\lambda_1^{(m)}\kappa_j\sum_{i=1}^{N}c_{ij}^{2(m)}} \tag{2.7}$$

$$= r^{(m)}\lambda_1^{(m)} \tag{2.8}$$

where, as in equation (2.5), $\sum_{i=1}^{N} c_{ij}^{2(m)}$ represents the variation present in the latent functions in the jth direction in iteration m. The full derivation of (2.6) can be found in Appendix A.

In (2.8), we observe that expression (2.7) can be reduced to $r^{(m)}\lambda_1^{(m)}$, where $0 < r^{(m)} < \frac{p+1+N}{p+1}$. The magnitude of $r^{(m)}$ is dependent on two values from the previous iteration of the sampler: 1) the curvature present in the latent functions measured by $\sum_{i=1}^{N} c_{ij}^{2(m)}, j = 3 \ldots p$, and 2) the last draw of the smoothing parameter, $\lambda_1^{(m)}$. This dependence can be described in the following way; for a given draw of $\lambda_1^{(m)}$, there exists some $K^{(m)}$ such that if $\lambda_1^{(m)} < K^{(m)}$, then $1 \leq r^{(m)} < \frac{p+1+N}{p+1}$ and as a result $E(\lambda_1^{(m+1)}) \geq \lambda_1^{(m)}$. Alternatively, if for this $\lambda_1^{(m)}$, $\lambda_1^{(m)} \geq K^{(m)}$, then $0 < r^{(m)} < 1$ and as a result $E(\lambda_1^{(m+1)}) < \lambda_1^{(m)}$. Furthermore, for any fixed $\lambda_1^{(m)}$, the threshhold, $K^{(m)}$, increases as the sums $\kappa_j \sum_{i=1}^{N} c_{ij}^{2(m)}$, $j = 3 \ldots p$, decrease. Thus, as we expect the values of $\kappa_j \sum_{i=1}^{N} c_{ij}^{2(m)}$, $j = 3 \ldots p$, to be small when the latent functions are smooth or the sample covariance function of the latent functions is singular, these conditions in general will result in a larger smoothing penalty and hence improved numerical stability. Alternatively, choosing λ_1 in an *ad hoc* manner often results in an unstable scale matrix that causes the sampler to fail.

In addition to the smoothing parameters associated with the scale matrix of the distribution on $\mathbf{\Sigma}_X$, the prior distributions for the mean function and the regression coefficient function also require parameters to provide regularization. Allowing unique smoothing parameters for each function provides flexible function-specific regularization. In selecting these parameters, we have found in practice that an automatic data driven approach to smoothing is not only practical but also results in a desirable amount of regularization. In Section 2.4, using medfly data, we compare the mean and regression coefficient functions estimated over a fixed range of smoothing parameters to the estimates determined by taking a stochastic approach to smoothing parameter selection. Figure 2.6 illustrates how our prior specifications for these smoothing parameters seem to allow for selecting parameters that are "just right." All of the prior dis-

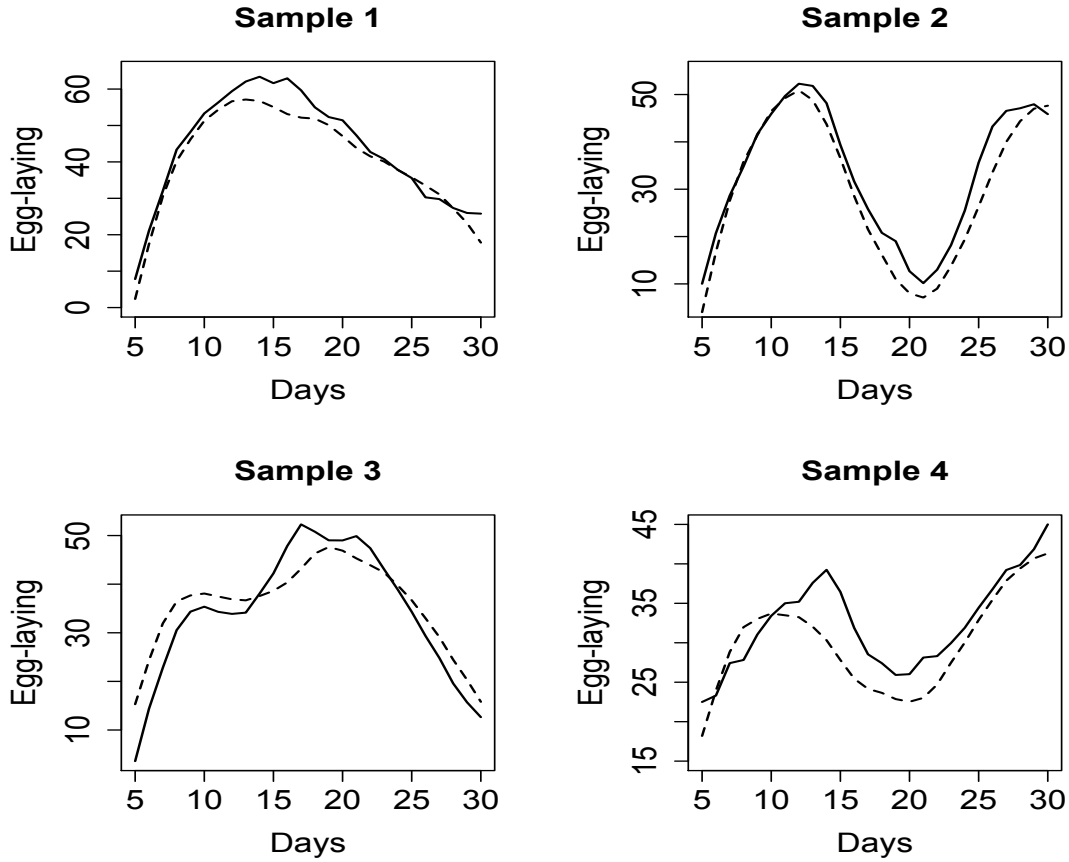

Figure 2.2: Comparison of the simulated and estimated functions. In each figure, the solid line is a simulated latent function and the dashed line is the estimate for that latent function using the model for estimating functional data described in Appendix A.2.1.

tributions utilized in the functional regression model can be found in Appendix A.2.2.

2.3 Simulation Results

In this section we present a simulation study to assess how well these models recover the latent functions $X_i(t)$ as well as the GP parameters $\mu(t)$ and $\Sigma_X(s,t)$ when these are known. Evaluations of 50 "latent" functions, $X_i(t), i = 1,\ldots,50$, at 26 time points, $\mathbf{t} = (5,6,\ldots,30)'$, are simulated from a Gaussian process to examine the estimation properties of the models described in this paper. $\mu(t)$ and $\Sigma_X(s,t)$ are set at the population mean and covariance function estimates from a subset of the medfly egg-laying data analyzed in Section 2.4. Observations, $Y_i(t_j)$, are then constructed such that $Y_i(t_j) = X_i(t_j) + \epsilon_i(t_j)$, with iid noise $\epsilon_i(t_j) \sim N(0,148)$, $i = 1,\ldots,50$, $j = 1,\ldots,26$. The latent curves and GP parameters were then reconstructed via the Gibbs sampler described in Appendix A.2.1.

We first note that the absolute difference of the actual and estimated variance of the iid zero mean noise is approximately .61. Thus, in this simulation analysis, the underlying noise process is accurately determined. In Figure 2.2, each of four illustrations contain a simulated function plotted with its estimated value under the assumed model. Comparing each simulated function to its estimated value, it can be seen that each estimated function tends to retain significant features of the corresponding "latent" function while smoothing out noisy behavior. Furthermore, plotting these sample functions with their estimated values demonstrates, as in classical approaches to imposing smoothing parameters, bias in the estimated functions tends to be greater in areas of high curvature. The fourth sample function, in Figure 2.2, particularly illustrates this phenomenon as roughness in the underlying function is considerably dampened in its corresponding estimate.

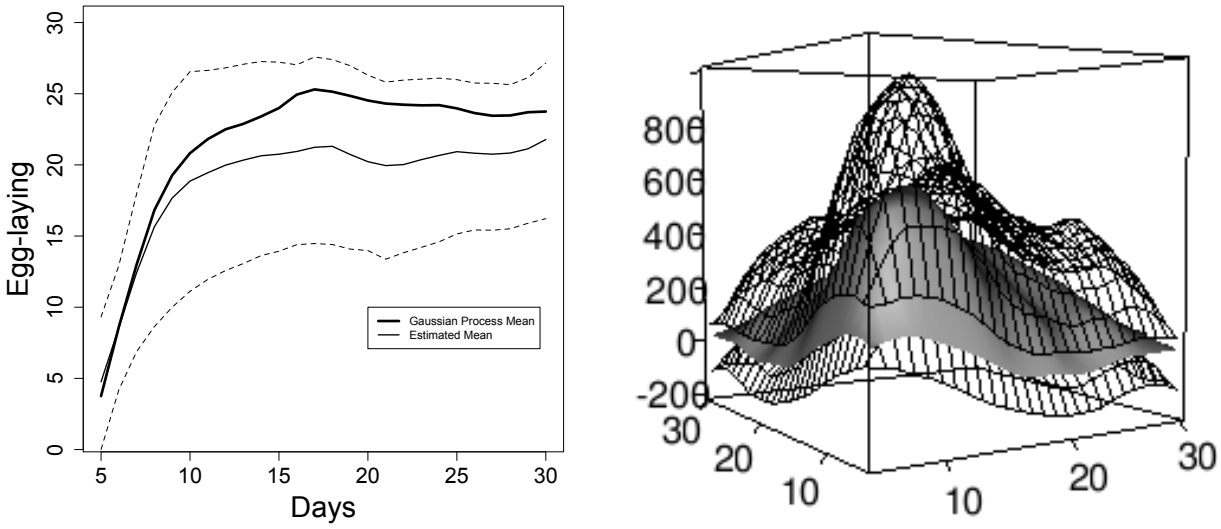

Figure 2.3: Comparison of the simulated and estimated mean and covariance functions. Ninety-five percent credible bands for the mean function used to simulate "latent" observations plotted with the estimated and actual mean function are plotted in the figure on the left. In the figure on the right, $\Sigma_X(s,t)$, $s,t \in (5,30)$, the covariance process used for simulation is the surface in gray while the wire mesh contains a 95% point wise credible area for the covariance function determined from the simulated observations.

In estimating the parameters of the Gaussian process that characterizes the latent functions, point-wise credible areas for the mean and covariance functions both encompass their corresponding population analogs. Figure 2.3 contains plots of these credible areas with the mean and covariance processes used for simulation.

2.4 Functional Regression Application: Medfly Fertility and Mortality

2.4.1 Medfly Data Analysis

A significant amount of literature has examined the relationship between medfly fertility and mortality. Here, we apply the functional regression method outlined in this paper, to again examine this relationship, primarily to illustrate the inherent properties of this model for estimation and other types of inference.

In this section, we re-analyze the medfly data of Müller and Stadtmüller [28] and apply the functional regression model (1.2) along with a GP model on covariates to examine the relationship between fertility and mortality. Additional information on the medfly egg-laying and mortality data analyzed in this example can be found in Müller and Stadtmüller [28] where the authors use a functional logistic regression model to classify 534 flies as long or short lived with egg-laying trajectories over the first 30 days of life as the predictor. In their model, the associated Bernoulli distribution models the probability of being a long-lived fly. Here, we ignore the first 4 days of egg-laying where egg-laying is frequently zero. The most significant finding of Müller's analysis of the relationship between medfly fertility and longevity indicates that high fertility later in life is associated with a longer lifespan. While our model uses a continuous response (total lifespan) instead of a binary response, as would be expected the estimated regression coefficient function under our model has a similar shape (but different scale) to that estimated by Müller.

Our covariates are represented by 26 time points which captures the period

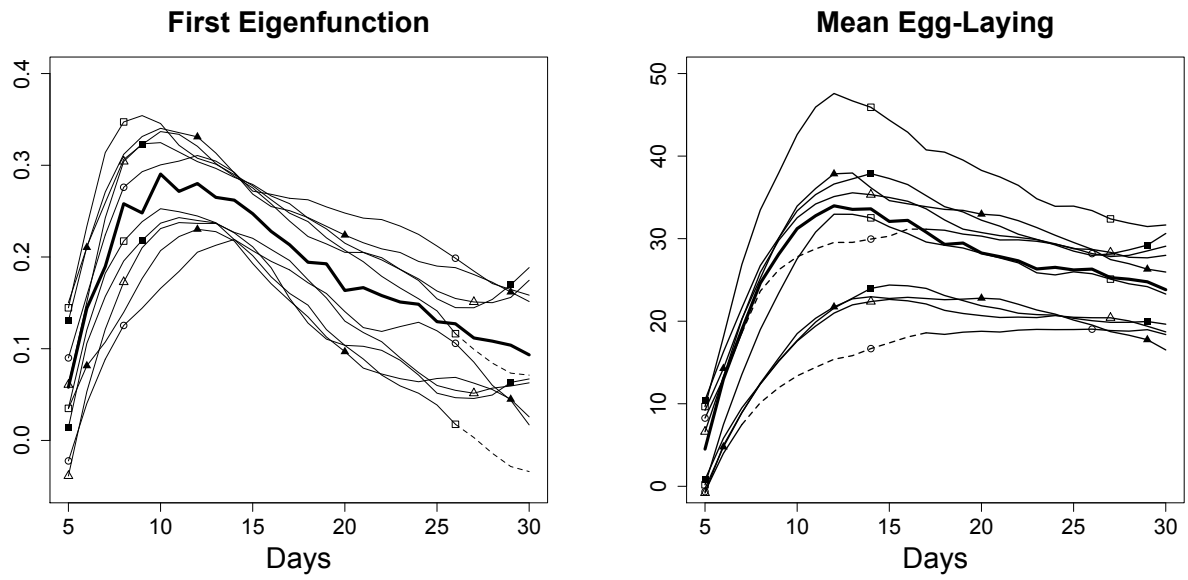

Figure 2.4: Estimated credible interval coverage of the first eigenfunction (left) and the mean function (right). The thick lines in each plot is the first eigenfunction or mean function determined from the full data set of 534 medflies. Plotted with the population means for the first eigenfunction and the mean function respectively are 95% point wise credible bands for the corresponding function determined from each of five subsets of the original data, where the upper and lower credible bands for a particular subset are designated by matching symbols. The dashed lines highlight portions of time where a credible interval that does not contain the population mean. Similar plots for the remaining 5 subsets of data can be found in Appendix A.3.

from initial fertility until fertility has dropped off significantly, but stops substantially before death. We use these data to illustrate the use of latent GP models for functional latent variables within a functional linear model (1.2). In our analysis, the response, z_i, is the total lifespan in days of fly i, for $i = 1, \ldots, 534$. The predictor, $X_i(t)$, $t \in [5, 30]$ is assumed to be a smooth biological process that generates the number of eggs laid in days 5 through 30. Observations $Y_i(t_j)$, $j = 1, \ldots, 26$, the total number of eggs laid by fly i on day t_j, are recorded to estimate the underlying biological processes, $X_i(t)$, for each of the 534 medflies.

Estimates of posterior distributions for all functional covariates, $X_i(t)$, the mean and covariance functions of these, $\mu(t)$ and $\Sigma_X(s,t)$, and the coefficient function, $\beta(t)$ are obtained by sampling from the joint posterior distribution. These samples provide both functional estimates and credible bands or surfaces for each unknown parameter. Furthermore, the sample of covariance surfaces allow variability in the covariance estimate to be expressed in terms of the function as a whole or through its eigenvectors and eigenvalues as desired.

All functional inference is dependent on smoothing parameters that are included as additional unknown parameters in the model as described in Section 2.2.2. Figure 2.6 illustrates the effect on the estimate of the mean and regression coefficient functions of selecting smoothing parameters that either over or under penalize curvature. The estimates of the mean and regression functions that are appropriately regularized are determined through allowing the smoothing parameters to be additional unknown parameters in the model.

For the purpose of examining the small-sample properties of this model, we have separated the data into ten subsets of 53 or 54 medflies that form a partition of the complete data set of 534 medflies. Organizing the data in this way allows us to consider the 534 medflies as the target population from which samples of sizes 53-54 are drawn. In particular, this allows us to conduct a "simulation" experiment to assess coverage in real-world data that may not correspond to our assumed model. We proceed by running the Gibbs sampler on each of the ten subsets to create ten posterior samples for all parameters. From these samples, we obtain ten parameter estimates and credible intervals for each parameter. This enables us to examine the credible interval coverage properties of this model empirically. We also compare results from runs where the covariance

function is assumed fixed versus our approach, where we define a distribution for the covariance function. This comparison demonstrates the effectiveness of our approach in estimating not only the uncertainty in the covariance function itself, but also the variability of all other estimated parameters inherited from it.

2.4.2 Covariance Estimation and Credible Interval Coverage

Here we use a "simulation" study to show that credible bands determined through the posterior samples provide approximately correct coverage. If we assume the population consists of the original data set of 534 flies, the 95% credible bands for the first eigenfunction determined from each of the 10 samples from this population ideally contain the population estimate of this eigenfunction 95% of the time. While estimating credible interval coverage from ten samples is a very rough measure of actual coverage, these 10 samples do give us some indication of credible interval behavior. In Figures 2.4 and A.1, credible interval bands for each of the ten samples have been plotted with the population estimate for the first eigenfunction. Of the ten 95% credible bands, all contain the estimated population eigenfunction except for the third sample where approximately 25% of the point wise credible band does not include the population estimate. So, roughly, for these ten samples, 97.5% of the credible intervals include the population estimate of the first eigenfunction. Thus, considering the small sample size, the empirical coverage of these credible intervals seems to be reasonable.

A similar analysis can be performed for the estimated mean function. Here, however, we compare the credible interval coverage under our model to the

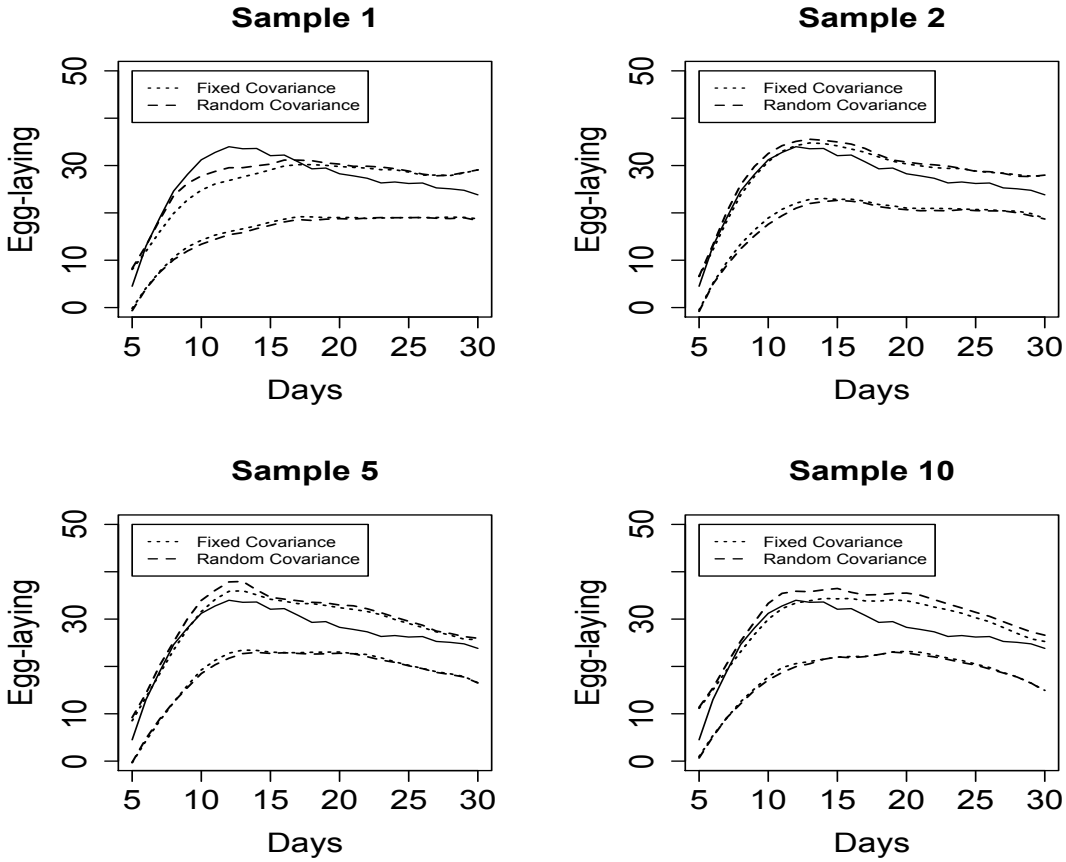

Figure 2.5: Comparison of credible band coverage under fixed and random covariance assumptions. The population mean function plotted with credible bands for each of four samples under fixed and stochastic covariance assumptions.

credible intervals obtained by fixing the covariance function in the GP model in advance instead of estimating it through the Gibbs sampler. This provides a comparison of our modeling approach to methods similar to those of Crainiceanu and Goldsmith [11]. Theoretically, we would expect inference methods that do not account for uncertainty in the estimate of the covariance function to underestimate variability in parameter estimates throughout the model. Our empirical analysis supports this theory.

For the estimates where the covariance function is modeled as known, the covariance function is fixed at the posterior sample mean of the covariance function from a previous run of the full GP model for that sample. Consequently, these estimates by definition reflect the best estimates for the covariance function for this model and *a priori* incorporate prior information and data. Since we utilize covariance function estimates from the full model as our "fixed" estimates, in the following analysis, the primary difference between the two models used for each sample is that in one the covariance function is considered stochastic and in the other the covariance function is considered known. As can be seen in Figures 2.4 and A.1, in the GP model where the covariance function is stochastic, the credible bands approximately cover the population mean 93% of the time. However, in the models where the covariance function has been held fixed, the resulting underestimation of variability is reflected in credible intervals that provide less coverage. Figure 2.5 presents a comparison of credible bands derived under the two different methods for four samples. Universally, the fixed covariance approach produces narrower credible intervals; over all ten samples they cover the population mean estimate 82% of the time. Hence, empirically, there is evidence credible bands obtained under a method where the covariance function or eigenfunctions are assumed known, but have actually been estimated prior to the inference procedure, do not provide adequate coverage.

2.4.3 Missing Data Results

One advantage of modeling missing data as latent parameters is that the resulting estimates of a latent function, at time points with missing observations,

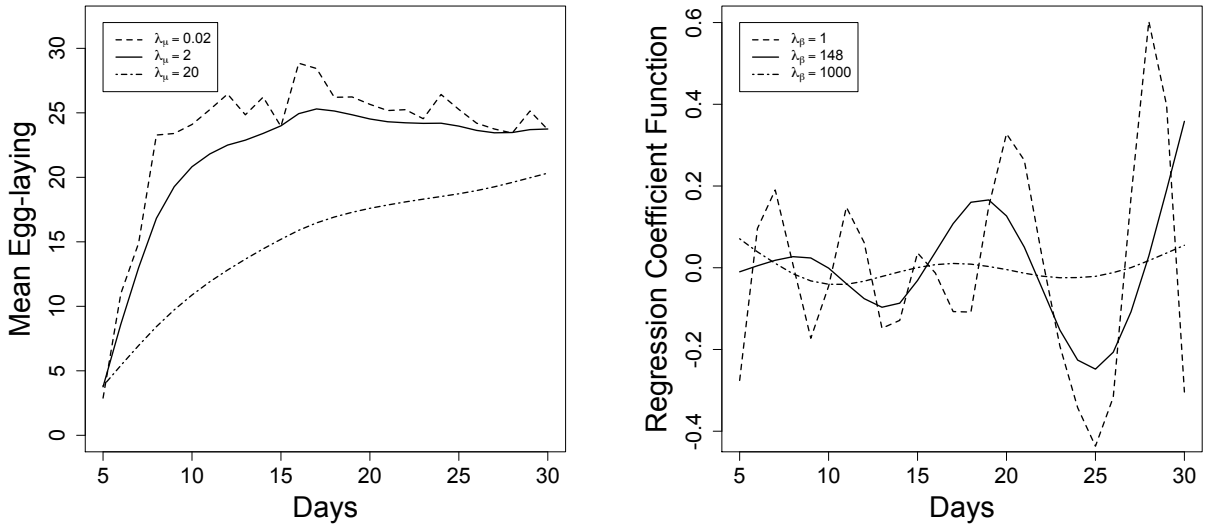

Figure 2.6: Results for a range of smoothing parameters. These plots highlight the sensitivity of parameter estimates to the choice of smoothing parameters. The solid lines are estimates with smoothing parameters chosen by the sampler.

draw on information from the mean and covariance process at those time points while also taking into account the smoothness of the process and neighboring observations. Here we will examine the effect of missing data under two scenarios. In the first, each function is missing blocks of data placed at random along the function. Here the remaining functions continue to provide information about the mean and covariance parameters. In the second scenario, the same blocks of data are missing for all functions, forcing our methods to rely on smoothing information to interpolate across the block.

Using sample one, observations are eliminated from the egg-laying data either consistently throughout the sample or in blocks of five adjacent observations, where the placement of the five observations varies from curve to curve. Where observations are systematically deleted from the data, all observations

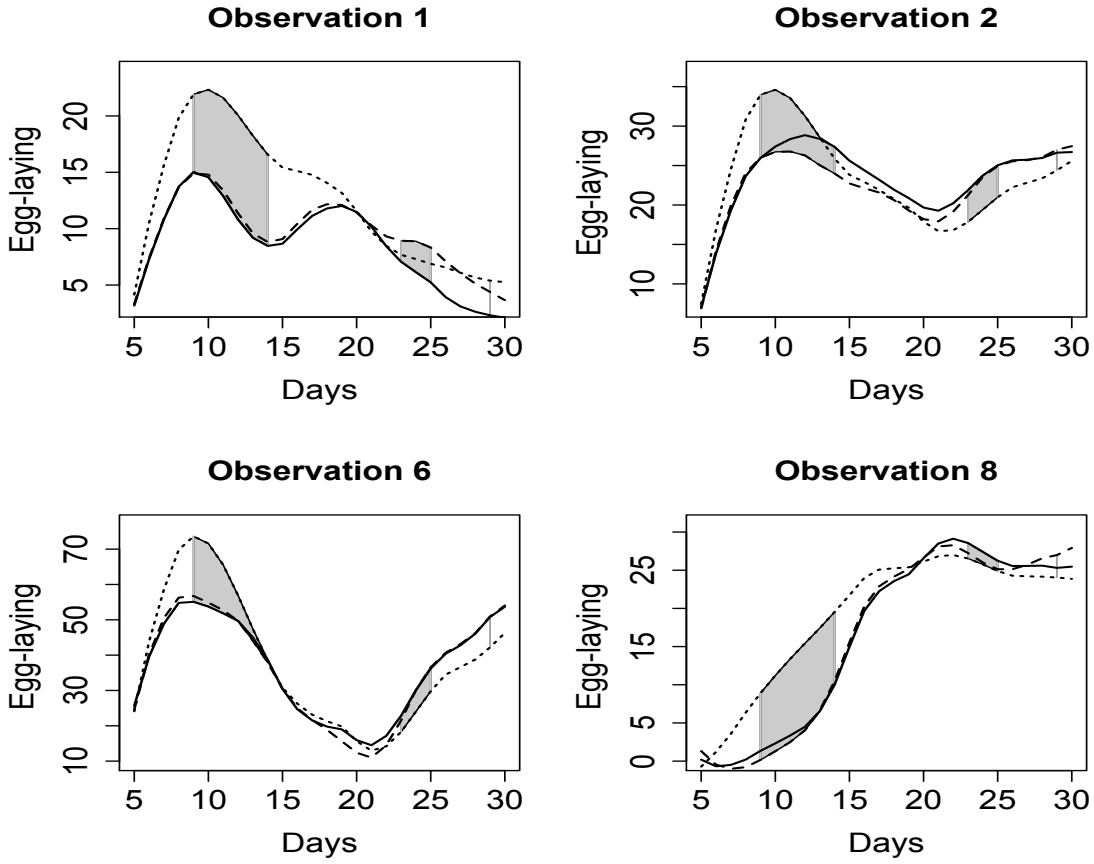

Figure 2.7: Function estimates using incomplete data. Each plot contains three estimates of latent functions from sample one. The solid lines are complete data estimates. The dashed lines represent estimates with data missing randomly in blocks. The dotted lines are estimates with data missing consistently in every observation at the time points corresponding to the shaded areas.

from days 9 through 14, 23 through 25, and day 29 are omitted.

The effect of missing observations is seen in the estimates of individual egg-laying trajectories. As shown in Figure 2.7, where the same blocks of data are omitted in every curve, estimates differ importantly from the curves that generated the data. This is not surprising as the estimate of a latent function in areas where no data are available rely only on the smoothing information in the prior

distribution and neighboring estimates of the latent function at observed time points. In contrast, trajectory estimates determined on the data set with random blocks of missing data look fairly similar to those estimated using the complete data. The close adherence of curves estimated with complete data to curves estimated with random deletions demonstrates how accurately this model estimates sections of missing data by using information from supporting curves that include observations at these time points.

CHAPTER 3

GAUSSIAN PROCESS MODELS FOR FUNCTIONAL DATA REGISTRATION, SMOOTHING, AND PREDICTION

3.1 Gaussian Process Models for Registration

The functional registration models proposed in this paper are foremost designed to extend and improve on the minimum eigenvalue registration criterion for continuous registration first introduced by Ramsay and Li [34] with additional extensions in Ramsay and Silverman [36]. A more detailed discussion of the registration model of Ramsey and Li and the extensions to this model presented in *Functional Data Analysis*, on which our model is based, can be found in Chapter 1. In concordance with this work, we will consider two functions perfectly registered if the variation between the two functions can be described entirely in terms of one functional direction. Our method of registration improves on Ramsay's Procrustes method by implicitly accounting for vertical shifts between registered functions, allowing the target curve to evolve throughout the registration procedure, and by providing a smoothing mechanism within the registration process. In Section 3.3, we will demonstrate how implicitly using the minimum eigenvalue criterion under these conditions provides a more complete curve registration. Our results are comparable to those of Srivistava, et.al. [43].

The primary advantage of our proposed registration model is that it provides a probabilistic framework in which new observations are considered. Using this framework, a new partially recorded observation can be registered to a corresponding piece of the target function, where the last registered time is

chosen over a range of times over which the target curve has been recorded. This partial registration provides an estimate of not only the registered partial function, but also the corresponding partial warping function. Using these estimates, the complete registered function, the complete warping function, and the complete unregistered function can be predicted. Details of the prediction model can be found in Section 3.4.

The theoretical basis for modeling functional data as Gaussian processes in a hierarchical Bayesian environment is established in Chapter 2. Here we continue this work by modeling each registered function, $X_i(h_i(t))$, $i = 1, \ldots, N$, as a Gaussian process such that

$$X_i(h_i(t)) \mid z_{0i}, z_{1i}, f(t) \sim GP(z_{0i} + z_{1i} f(t), \gamma_R^{-1}\Sigma(s,t)) \quad s,t \in \mathcal{T} \tag{3.1}$$

The above covariance function, $\gamma_R^{-1}\Sigma(s,t)$, penalizes all variance from a scaling and vertical shifting of the target function, $f(t)$. In these models we will define γ_R as a registration parameter that determines the severity of this penalty. This registration parameter is balanced by a penalty on the warping functions, $h_i(t), i, \ldots, N$ that penalizes distance from the identity warping as well as smoothness in the warping function.

Given a sample of unregistered functions, $X_i(t)$, $i = 1, \ldots, N$, defined over the interval $\mathcal{T} = [t_1, t_p]$, we are interested in estimating the warping functions, $h_i(t)$, the shifting and scaling parameters, z_{0i} and z_{1i}, the target curve, $f(t)$ and the registered functions.

For now, we will assume the functions are recorded without noise. If the functions are recorded with noise, it is common practice in the current literature to first perform a pre-processing smoothing step. An undesirable result of this pre-processing step is that the subsequent inference procedure is unable to

capture the extra variability associated with the smoothing process. In Section 3.5, we will show how our model can be modified to both smooth and register functions.

Inference is accomplished through a Bayesian hierarchical model. The distributional assumptions and prior specifications for this model are,

$$
\begin{aligned}
X_i(h_i(t)) \mid z_{0i}, z_{1i}, f(t) &\sim GP(z_{0i} + z_{1i} f(t), \gamma_R^{-1}\Sigma(s,t)) \quad s,t \in \mathcal{T} \quad i = 1,\ldots,N \\
h_i(t) &= t_1 + \int_{t_1}^{t} e^{w_i(s)} ds \quad t \in \mathcal{T} \quad i = 1,\ldots,N \\
w_i(t) &\propto GP(0, \gamma_w^{-1}\Sigma(s,t) + \lambda_w^{-1} P_w(s,t)) \mathbb{1}\{t_1 + \int_{t_1}^{t_p} e^{w_i(s)} ds = t_p\} \\
&\quad s,t \in \mathcal{T} \quad i = 1,\ldots,N \\
z_{0i} \mid \sigma_{z0}^2 &\sim N(0, \sigma_{z0}^2) \quad i = 1,\ldots,(N-1) \quad z_{0N} = -\sum_{i=1}^{N-1} z_{0i} \\
\sigma_{z0}^2 &\sim IG(a,b) \\
z_{1i} \mid \sigma_{z1}^2 &\sim N(1, \sigma_{z1}^2) \quad i = 1,\ldots,N \qquad (3.2) \\
\sigma_{z1}^2 &\sim IG(a,b) \qquad (3.3) \\
f(t) \mid \eta_f, \lambda_f &\sim GP(0, \Sigma_f(s,t)) \quad s,t \in \mathcal{T} \\
\Sigma_f(s,t) &= \eta_f^{-1} P_1(s,t) + \lambda_f^{-1} P_2(s,t) \\
\eta_f &\sim G(c,d) \\
\lambda_f &\sim G(c,d) \qquad (3.4)
\end{aligned}
$$

Note, for this model, the distribution on $\mathbf{z}_1 = (z_{11} \ldots z_{1N})'$ can be replaced by a Dirichlet distribution on $\mathbf{z}_1/N$. The result is a slightly more complicated model that has the nice effect of scaling the target function to the empirical mean of the estimated registered functions. Priors (3.2) and (3.3) would then be omitted.

Given the above model, in practice we will proceed by using finite approx-

imations to each functional distribution. In Section 2.1.2, we establish some theoretical properties of these types of approximations. The following finite-dimensional distributions are used in the final model in lieu of their infinite dimensional counterparts above.

$$\begin{aligned}
\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f} &\sim N_p(z_{0i}\mathbf{1} + z_{1i}\mathbf{f}, \gamma_R^{-1}\boldsymbol{\Sigma}) \quad i = 1, \ldots, N && (3.5)\\
\mathbf{h}_i(t_j) &= t_1 + \sum_{k=2}^{j}(t_k - t_{k-1})e^{w_i(t_{k-1})} \quad i = 1, \ldots, N \quad j = 1, \ldots, p \\
\mathbf{w}_i &\propto N_{p-1}(\mathbf{0}, \gamma_w^{-1}\boldsymbol{\Sigma} + \lambda_w^{-1}\mathbf{P}_w)\mathbb{1}\{t_1 + \sum_{k=2}^{p}(t_k - t_{k-1})e^{w_i(t_{k-1})} = t_p\} && (3.6)\\
&\quad i = 1, \ldots, N \\
\mathbf{f} \mid \eta_f, \lambda_f &\sim N_p(0, \boldsymbol{\Sigma}_f) \\
\boldsymbol{\Sigma}_f &= \eta_f^{-1}\mathbf{P}_1 + \lambda_f^{-1}\mathbf{P}_2
\end{aligned}$$

The sum of the matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ form a basis for R^p. The matrix, $\mathbf{P}_2 = (\mathbf{L}'\mathbf{L})^-$, where $\mathbf{L}$ is a second finite difference matrix of rank $p-2$. The eigenvectors of $\mathbf{L}$ approximately span the space of all functions that are not constant or linear. The matrix $\mathbf{P}_1 = (\mathbf{B}'\mathbf{B})^-$ where $\mathbf{B}$ is an approximated squared constraint operator so that if B is the constraint operator for any function, $X(t)$, $BX(t) = [X(0), X'(0)]$. By this definition, $\mathbf{B}$ is an approximate basis for constant and linear functions. $\mathbf{B}$ and $\mathbf{L}$ together span all of R^p and the eigenvectors of $\mathbf{B}$ are orthogonal to the eigenvectors of $\mathbf{L}$. The matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ serve two purposes in this registration model. When considered together, they define the matrix $\boldsymbol{\Sigma} = \mathbf{P}_1 + \mathbf{P}_2$ that penalizes any variation from a given mean function. Considered separately, as in $\boldsymbol{\Sigma}_f$, $\mathbf{P}_2$ provides a penalty on roughness with associated smoothing parameter λ_f. $\mathbf{P}_1$ is only necessary in $\boldsymbol{\Sigma}_f$ to assure this covariance matrix is non-singular and consequently η_f is only large enough to provide sta-

bility in this matrix. Here, η_f and λ_f are estimated within the model. For a more extensive discussion on selecting smoothing parameters in this way, see Section 3.5. Finally, the matrix $\mathbf{P}_w$ is only present to provide extra smoothness in the warping functions if necessary. It can be denoted as either a squared first or second derivative penalty on the base functions and is sometimes excluded altogether. Exact definitions of these covariance matrices can be found in equation (2.3).

In this paper, we will refer to the functions, $w_i(t)$, $t \in \mathcal{T}$, from which the warping functions, $h_i(t)$, $t \in \mathcal{T}$, are derived, as the base functions. The base functions are non-parametrically specified for optimal registration. We, however, impose the following restrictions on the warping functions:

1. $h(t_1) = t_1$
2. $h(t_p) = t_p$
3. if $t_k > t_j$, then $h(t_k) > h(t_j)$ for all $t_k, t_j \in \mathcal{T}$

Restrictions (1) and (3) are built into the definition of $h_i(t)$. Restriction (2) is imposed through the characteristic function in the expression for the prior defined for each base function, $w_i(t)$. Note that $w_i(t) = 0$ corresponds to the identity warping, $h_i(t) = t$. An important feature of our definition of the warping functions is that it defines an identifiable relationship between $h_i(t_j)$ and $w_i(t)$ which is necessary for predicting future outcomes of curves only partially observed. In Section 3.4 is a more thorough discussion of the prediction model.

Effectively two penalties are defined on the warping functions in this model. In the specification of the covariance function for the registered functions,

$\gamma_R^{-1}\Sigma(s,t)$, penalizes variation beyond a scaling and vertical shift from the target function. Here we will define γ_R as a registration parameter that controls the extent of function registration. The second penalty controls the amount the warping functions can deviate from the identity warping. This penalty is assessed through the covariance function, $\gamma_w^{-1}\Sigma(s,t)$, in the prior defined on the base functions, $w_i(t)$, $i,\ldots,N$. The warping parameter, γ_w, controls the extent the warping functions are allowed to deviate from the identity warping. Sometimes it is also helpful to penalize either the first or second derivative of the base functions to assure there is smoothness in the transformation. For this purpose, an additional penalty is introduced in the covariance parameter of the base functions, $\lambda_w^{-1}P_w(s,t)$. However, this penalty is not always necessary, and for some analyses $\lambda_w \approx 0$ is appropriate. For a given statistical analysis, γ_R, γ_w, and λ_w can be adjusted to allow for a desirable amount of warping. This model can also be adapted to allow for function specific warping penalties. In Section 3.4.2, we will give an example where function specific penalties for the base functions have been utilized to preserve significant covariance relationships in the estimated registered functions.

While warping procedures perform best when all functions are a scaling of and vertical shift from a target function, it is common in practice for unregistered datasets to include functions that vary significantly in other directions. In this model, a large registration penalty in comparison to the penalty on warping will result in registered functions that no longer retain significant features in the data. Alternatively, a registration parameter that is too small will not properly align features. The desired values of these parameters can be best determined through the adapted variational Bayes algorithm described in Section 3.2.1. Once determined, γ_R, γ_w, and λ_w are fixed and can be used with the

adapted variational Bayes estimates to initialize an MCMC sampler. Alternatively, we will show in Section 3.3.2 that differences in the estimated parameters obtained through adapted variational Bayes and MCMC sampling tend to be small, and estimation via adapted variational Bayes alone is likely sufficient for many inferential procedures.

3.2 Variational Approximation for Bayesian Registration

3.2.1 Adapted Variational Bayes

For this hierarchical Bayesian model, it is appropriate to use Markov Chain Monte Carlo (MCMC) methods to sample from the joint posterior distribution of all unknown parameters. The advantage of using a MCMC model for registration is that it is clear upon inspection whether the chain has been run long enough to provide good estimates. However, for most applications, the dimensionality of the parameter space will require exceptionally long chains that are impractical and expensive to obtain. Consequently, we suggest a variational Bayes alternative to MCMC to at the very least obtain good starting values for a MCMC sampler.

The variational Bayes procedure described here is based on the variational methods proposed by Omerod and Wand [30] and Bishop [2]. Their proposed method optimizes a lower bound of the marginal likelihood which results in finding an approximate joint posterior density that has the smallest Kullback-Leibler (KL) distance from the true joint posterior density. In Goldsmith et. al. [16] the authors utilize variational Bayes for a functional regression model and

provide a clear explanation of the procedure and how convergence of the parameter estimates is monitored.

It is important to note that the form of variational Bayes from which our estimation procedure is derived requires the approximate posterior distribution to factor over a partition of the set of parameters. Thus, effectively, the subsets of the partition are assumed to be independent. Additionally, the prior distributions in these models are assumed to be conditionally conjugate to the likelihood function. In minimizing the KL distance between the approximate and true posterior distribution, parameters are updated by an optimization method that requires an approximate posterior density that not only factors but factors into components of known parametric forms. Suppose, $q(\boldsymbol{\theta})$ is the approximated posterior joint distribution. Then for some partition of $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d\}$, $q(\boldsymbol{\theta}) = \prod_{k=1}^{d} q_k(\boldsymbol{\theta}_k)$, where each distribution q_k is of a known parametric form.

In our model, the Gaussian process priors for the base functions, $w_i(t)$, $i = 1, \ldots, N$, are not conditionally conjugate to the likelihood function. Therefore, the traditional variational Bayes optimization method does not apply directly since $q_k(\mathbf{w}_i)$, $i = 1, \ldots, N$ are not known parametric distributions.

3.2.2 Adapted Variational Bayes For Functional Data

As mentioned above, the prior distributions on the approximated base functions, $\mathbf{w}_i$, $i = 1, \ldots, N$, are not conditionally conjugate with the data distribution, $\prod_{i=1}^{N} f(\mathbf{X}_i(\mathbf{h}_i) \mid \boldsymbol{\theta})$. However, if we define $\boldsymbol{\theta}_k = \mathbf{w}_k$ for $k = 1, \ldots, N$ so that, $\boldsymbol{\theta} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N, \boldsymbol{\theta}_{N+1}, \ldots, \boldsymbol{\theta}_d\}$, for $k = \{(N+1), \ldots, d\}$, $q_k(\boldsymbol{\theta}_k)$ are known parametric distributions that can be estimated using the standard variational Bayes algo-

rithm. The traditional variational Bayes procedure can be reduced to the following:

1. Initialize $\boldsymbol{\theta}$
2. For each iteration, m, and each k, $k = 1, \ldots, d$, update q_k so that $q_k^{(m)} \propto exp[E_{(\theta_{-k})}(log\ f(\boldsymbol{\theta}_k \mid rest)]$, where the expectation is taken with respect to the distributions $q_j^{(m-1)}(\boldsymbol{\theta}_j)$, $j = 1 \ldots d$, $j \neq k$
3. Repeat step (2) until the desired convergence criterion is met

Here the notation, $E_{(\theta_{-k})}$, denotes the expected value over all parameters except θ_k. In the next section, we will drop the subscript k, and $E_{(\theta_{-\theta_k})}$ will represent the expectation over all parameters except for θ_k (e.g. $E_{(\theta_{-\eta_f})}$ will represent the expectation taken over all parameters except for η_f).

We will add an extra step to this algorithm that iteratively updates an estimate for each approximated function $\mathbf{w}_i, i = 1, \ldots, N$ so that the new algorithm is:

1. Initialize $\boldsymbol{\theta}$
2. For each iteration, m, and each k, $k = 1, \ldots, N$, update the estimate for $\mathbf{w}_k$ so that $\mathbf{w}_k^{(m)} = \sup_{\mathbf{w}_k} q_k(\mathbf{w}_k \mid \boldsymbol{\theta}_j^{(m-1)}, j = (N+1), \ldots, d)$
3. For each iteration, m, and each k, $k = (N+1), \ldots, d$, update q_k so that $q_k^{(m)} \propto exp[E_{(\boldsymbol{\theta}_{-k})}(log\ f(\boldsymbol{\theta}_k \mid rest))]$, where the expectation is taken with respect to the distributions $q_j^{(m-1)}(\boldsymbol{\theta}_j)$, $j = (N+1), \ldots, d$, $j \neq k$
4. Repeat steps (2) and (3) until the desired convergence criterion is met

Effectively, our adapted variational Bayes estimation procedure first maximizes the likelihood function in the warping functions with all other parameters fixed at their values determined by the previous iteration. We then consider the updated likelihood function where not only the data, but also the warping functions are assumed known and are fixed at the values determined in step 2. In step 3, using this new likelihood function, the rest of the parameters are updated by a traditional variational Bayes iteration. Under these assumptions, convergence is guaranteed and the estimation procedure can be monitored under the same criterion established by traditional variational Bayes.

However, convergence is not guaranteed to a global maximum, and in practice it is sometimes necessary to adjust the registration and warping penalties as the functions become registered. An unregistered function that requires a substantial amount of warping can cause convergence to a local maximum due to the small penalty on warping. The flexibility in warping allowed by this small penalty can cause the function to deform rather than register. This can be remedied in two ways. The first option is to perform a simple initial warping for this function that prevents the optimization from falling into a local mode. The second option is to adjust the registration and warping parameters over time. Initially a stronger warping penalty is employed to prevent function deformation. Then, as the functions register, the warping penalty can be reduced to allow for a more complete registration. When initializing an MCMC sampler, the final penalties on warping and registration from the adapted variational Bayes algorithm should be used. In Section 3.2.3 below is a more detailed discussion of the convergence properties of this model.

Section 3.3 provides several examples that illustrate how allowing the target function to be estimated within the model results in a more complete functional registration in comparison to the Procrustes method. However, the Gaussian process model does not constrain the timing of a feature in the target function to occur at the average time of the corresponding unregistered features. Although the model for the $w_i(t)$ is centered on zero, it is still possible that the average of the estimated warped time points $\overline{h_\cdot(t_1)},\ldots,\overline{h_\cdot(t_p)}$ does not correspond to the original time-points. Shifting these by an additional registration so that the warped times average to the original time does not affect our prediction model, but it will then allow an explicit comparison of $h_i(t_j)$ to t_j to tell us whether the process is running ahead or behind "standard" time.

Consider the following:

1. If the functions are registered so that registered features occur at the average time that they appear in the unregistered sample, for all t, $t \in \{t_1,\ldots,t_p\}$, the average warping at that time point, $\overline{h_\cdot(t)} = \frac{1}{N}\sum_{i=1}^{N} h_i(t)$, is the identity. Over all observed time points this implies $\overline{\mathbf{h}}_\cdot = (\overline{h_\cdot(t_1)} = t_1,\ldots,\overline{h_\cdot(t_p)} = t_p)'$.
2. Generally, after the registration process, this property will not hold, and instead $\overline{\mathbf{h}}_\cdot = (\overline{h_\cdot(t_1)} = \tilde{t}_1,\ldots,\overline{h_\cdot(t_p)} = \tilde{t}_p)'$ where $t_j \neq \tilde{t}_j$ for at least one $j \in \{1,\ldots,p\}$.
3. The goal is to shift the functions so that these average warpings correspond to the correct registered times, i.e. $\overline{\tilde{\mathbf{h}}}_\cdot = (\overline{h_\cdot(\tilde{t}_1)} = \tilde{t}_1,\ldots,\overline{h_\cdot(\tilde{t}_p)} = \tilde{t}_p)'$
4. (3) implies that after the initial registration, we have the correct registered function values over the new set of times, $\tilde{\mathbf{t}} = (\tilde{t}_1,\ldots,\tilde{t}_p)'$, i.e. we have estimates of $\mathbf{X}_i(\tilde{\mathbf{h}}_i) = (X_i(h_i(\tilde{t}_1)),\ldots,X_i(h_i(\tilde{t}_p)))'$, for all $i = 1,\ldots,N$.

5. If it is desired, the estimated values of the registered functions at the original time points, $\mathbf{t}$, can be obtained by interpolating values between the new set of time points, $\tilde{\mathbf{t}}$.

For example, after the initial registration process, suppose $\overline{h_{\cdot}(2)} = 2.25$ and $\overline{h_{\cdot}(3)} = 3.1$, where 2 and 3 are in the set of original time points. We can alter registered time so that $\overline{h_{\cdot}(2.25)} = 2.25$ and $\overline{h_{\cdot}(3.1)} = 3.1$, as desired. Using the notation above, this implies $\{2.25, 3.1\} \subset \tilde{\mathbf{t}}$ and for all $i, i = 1 \ldots, N$, from the initial registration, we estimated the following values, $X_i(h_i(2.25))$ and $X_i(h_i(3.1))$. From these we can estimate $X_i(h_i(3))$ by interpolating the values $X_i(h_i(2.25))$ and $X_i(h_i(3.1))$.

Note: Srivistava, et.al. [43] use a similar "correction" to determine their target function.

3.2.3 Convergence Criterion

The objective of the traditional variational Bayes algorithm is to find an approximate joint posterior distribution, given certain independence constraints, that has the minimum Kullback-Leibler distance from the true joint posterior distribution. It can be shown that minimizing the distance between the approximate posterior distribution, $q(\boldsymbol{\theta})$, and true joint posterior distribution, $f(\boldsymbol{\theta} \mid data = \mathbf{X})$, is equivalent to maximizing a q-specific lower bound on the marginal distribution of $\mathbf{X}$ defined as

$$f(\mathbf{X}; q) := exp \int q(\boldsymbol{\theta}) log \left\{ \frac{f(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta}$$

This result is based on the definition of the Kullback-Leibler distance, Kullback and Leibler [23]; details can be found in Goldsmith et. al. [16].

This algorithm is guaranteed to converge. Furthermore, when $q(\boldsymbol{\theta})$ is defined as a product of exponential distributions, the algorithm converges to a global maximum of $f(\mathbf{X}; q)$. Traditional variational Bayes monitors $log[f(\mathbf{X}; q)]$ until changes in this value are below a set threshold. Here we will show when the functional data are assumed to be observed without noise that our proposed adapted variational Bayes algorithm also is guaranteed to converge, and convergence can be monitored in a similar fashion as traditional variational Bayes.

The natural log of the lower bound of the marginal distribution of $\mathbf{X}$ can be expressed as

$$\begin{aligned} log\ f(\mathbf{X}; q) &= \int q(\boldsymbol{\theta}) log \left\{ \frac{f(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta} \\ &= E_{q(\boldsymbol{\theta})}[log[f(\mathbf{X}, \boldsymbol{\theta})] - log[q(\boldsymbol{\theta})]] \end{aligned}$$

As above, for this model, define $\boldsymbol{\theta} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N, \boldsymbol{\theta}_{N+1}, \ldots, \boldsymbol{\theta}_d\} := \{\mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}}\}$. The adapted variational Bayes algorithm above, alternates between: 1) maximizing $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log[f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}})]]$ in $\mathbf{w}$, and 2) fixing $\mathbf{w}$ at the value determined by (1) and using traditional variational Bayes to update a lower bound of the marginal distribution of $\mathbf{X}$ and $\mathbf{w}$,

$$\begin{aligned} log\ f(\mathbf{X}, \mathbf{w}; q) &= \int q(\boldsymbol{\theta}_{-\mathbf{w}}) log \left\{ \frac{f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}})}{q(\boldsymbol{\theta}_{-\mathbf{w}})} \right\} d\boldsymbol{\theta}_{-\mathbf{w}} \\ &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log[f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}})] - log[q(\boldsymbol{\theta}_{-\mathbf{w}})]] \end{aligned}$$

For each iteration, m of our adapted variational Bayes algorithm,

$$\begin{aligned} log\ f(\mathbf{X}, \mathbf{w}^{(m)}; q^{(m)}) &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log[f(\mathbf{X}, \mathbf{w}^{(m)}, \boldsymbol{\theta}_{-\mathbf{w}}^{(m)})] - log[q(\boldsymbol{\theta}_{-\mathbf{w}}^{(m)})]] \\ &\leq E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log[f(\mathbf{X}, \mathbf{w}^{(m+1)}, \boldsymbol{\theta}_{-\mathbf{w}}^{(m)})] - log[q(\boldsymbol{\theta}_{-\mathbf{w}}^{(m)})]] \quad (3.7) \\ &\leq E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log[f(\mathbf{X}, \mathbf{w}^{(m+1)}, \boldsymbol{\theta}_{-\mathbf{w}}^{(m+1)})] - log[q(\boldsymbol{\theta}_{-\mathbf{w}}^{(m+1)})]] \quad (3.8) \\ &= log\ f(\mathbf{X}, \mathbf{w}^{(m+1)}; q^{(m+1)}) \end{aligned}$$

The inequality in (3.7) is guaranteed by step 1 of the adapted variational Bayes algorithm, and the inequality in (3.8) is the result of using the traditional variational Bayes algorithm with $\mathbf{w}$ considered known. Consequently, as in traditional variational Bayes, this monotonic increasing sequence is guaranteed to converge. However, convergence to a global maximum is not guaranteed. In practice, occasionally maximizing $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log[f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}})]]$ in $\mathbf{w}$ results in a local optimization for some $\mathbf{w}_i$, $i \in \{1 \ldots N\}$. For a well-defined registration problem, this is easily discovered by inspection and can be overcome by initializing the troublesome $\mathbf{w}_i$ to a value that more closely represents the true warping.

The lower bound of the marginal distribution of $\mathbf{X}$ and $\mathbf{w}$ can be monitored until changes in this function are under some threshold. The specific form of this function can be found in Appendix B.4.

3.3 Comparison to Current Methods

3.3.1 Comparison to Other Registration Procedures

Our registration criterion minimizes all variation in the warped functions that is not in the direction of the target function (allowing for vertical shifts). In this

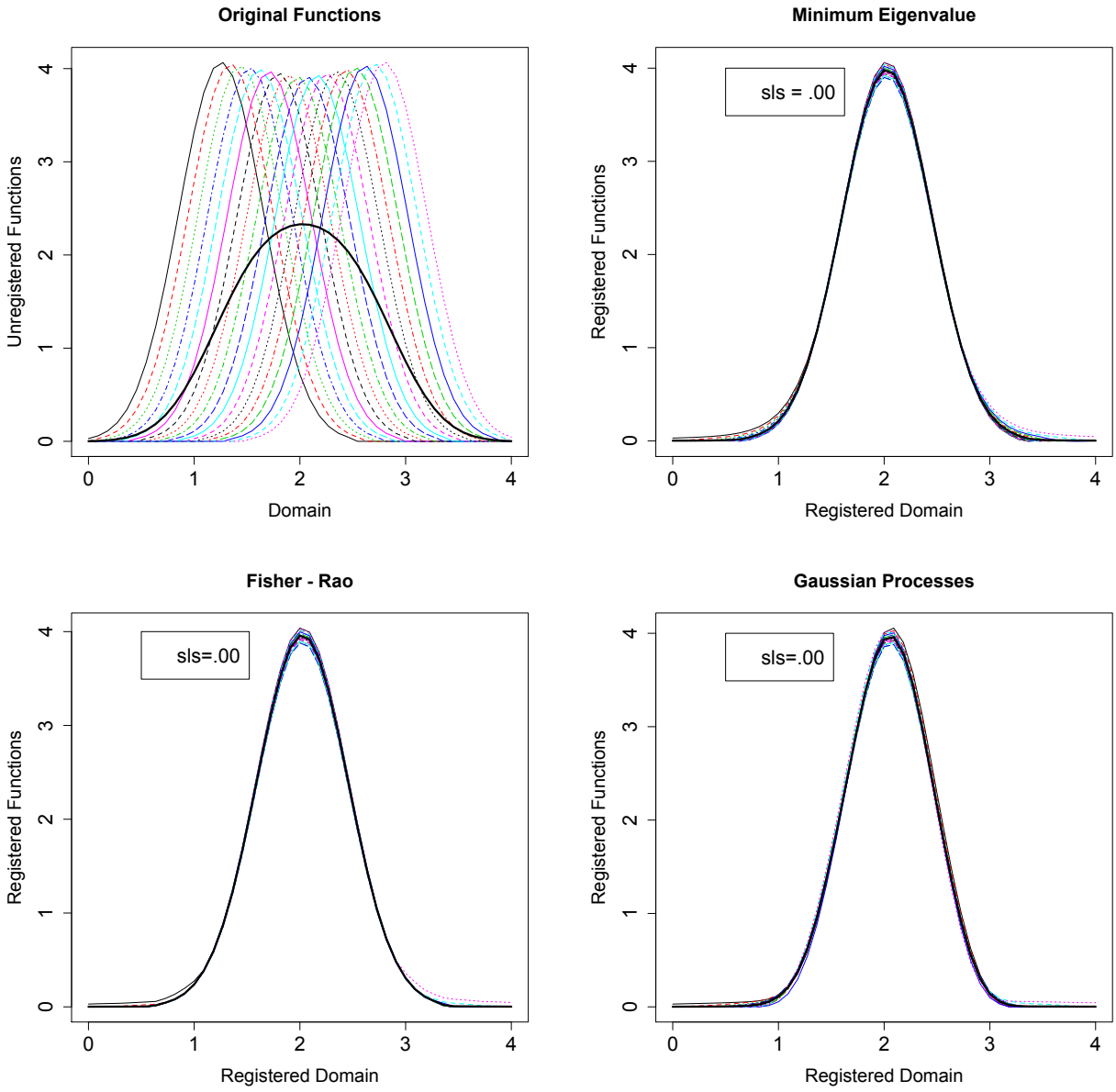

Figure 3.1: Simulated Data Set 1. **Top Left** Unregistered functions. **Top Right** Registered functions using the minimum eigenvalue criteria (R package 'fda'). **Lower Left** Functions registered by F-R (R package 'fdasrvf'). **Lower Right** Functions registered by the GP model.

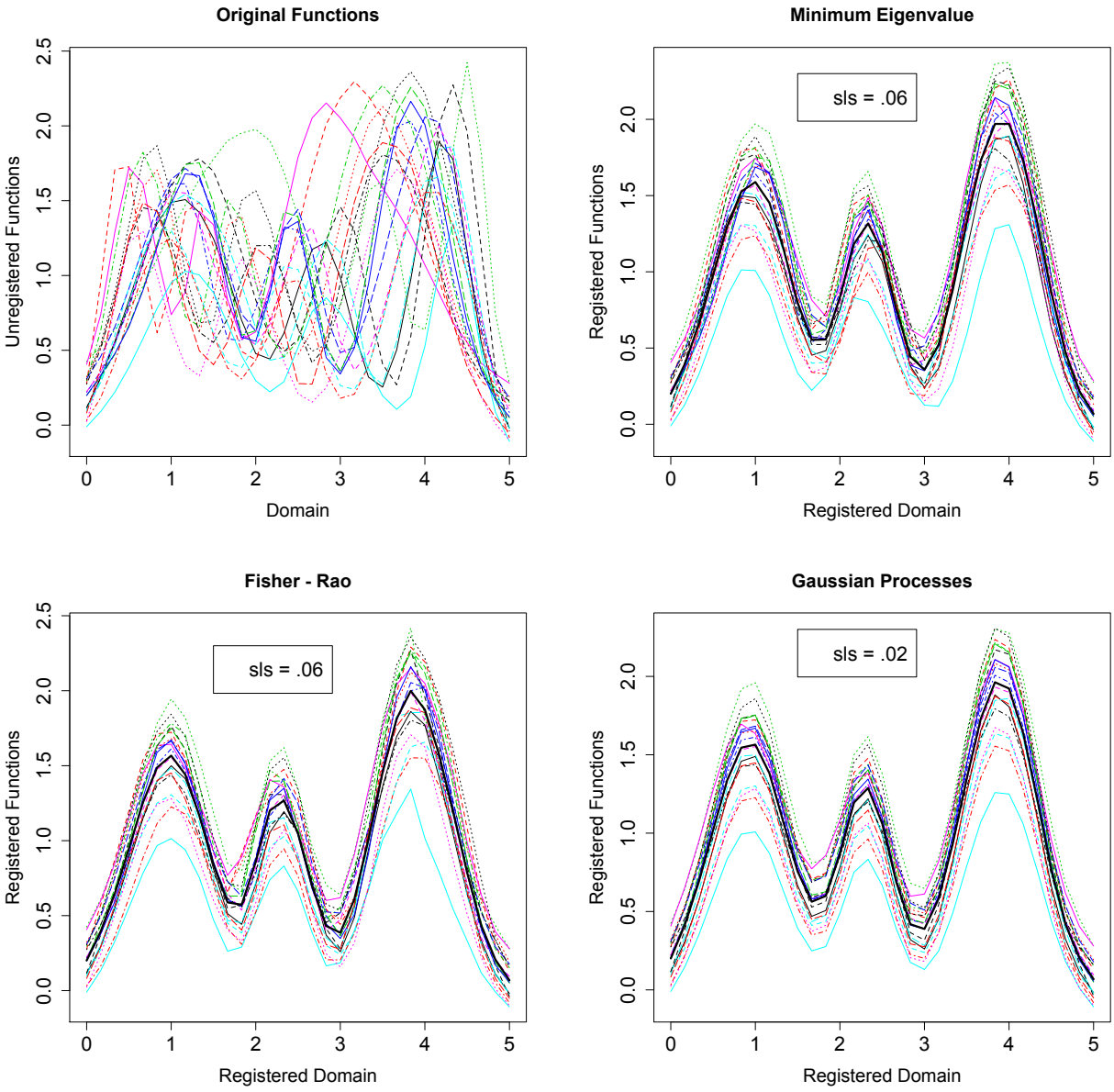

Figure 3.2: Simulated Data Set 2. **Top Left** Unregistered functions. **Top Right** Registered functions using the minimum eigenvalue criteria (R package 'fda'). **Lower Left** Functions registered by F-R (R package 'fdasrvf'). **Lower Right** Functions registered by the GP model.

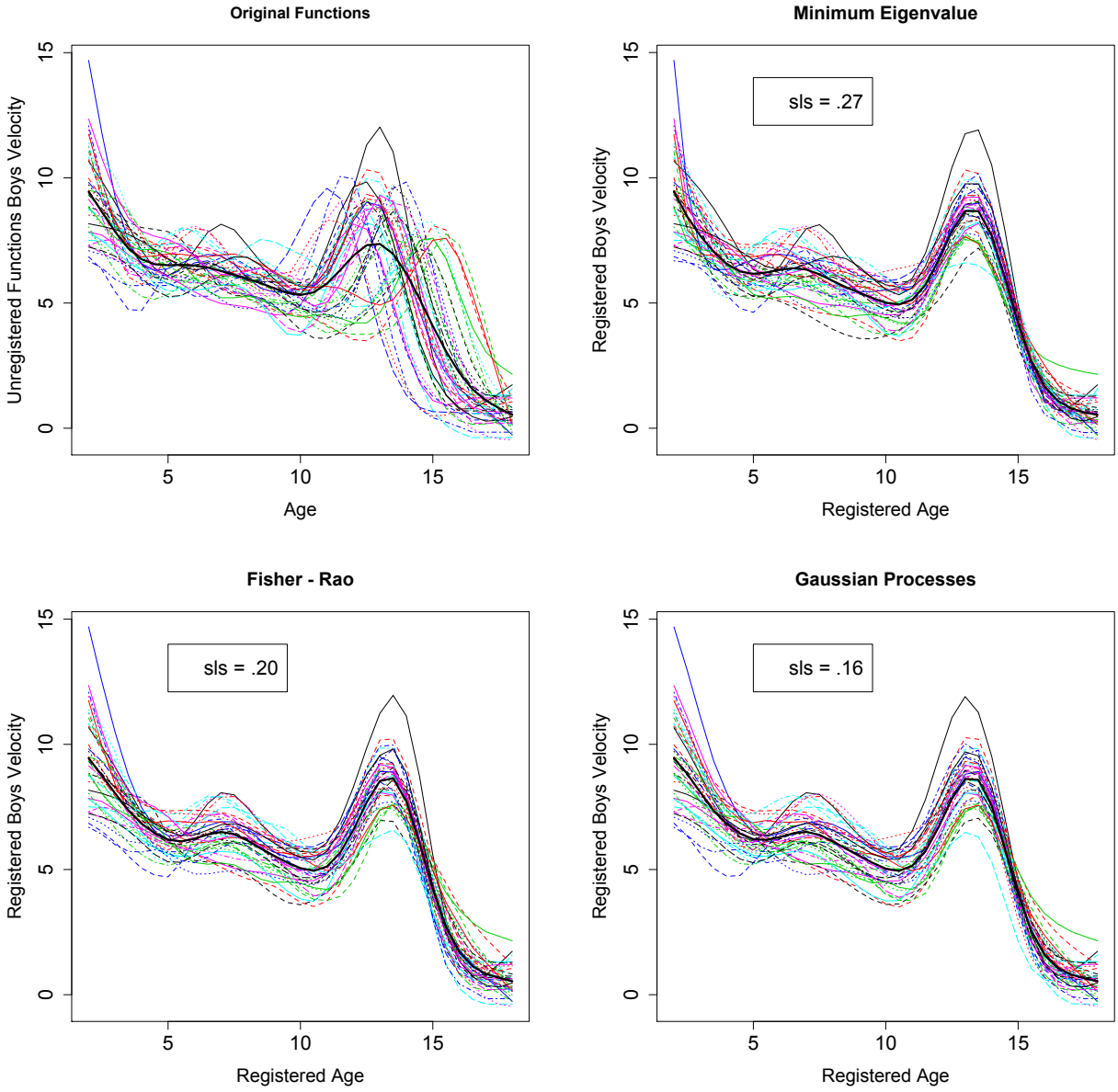

Figure 3.3: Registered Boys Growth Velocity. **Top Left** Original unregistered boys velocity data functions. **Top Right** Boys velocity functions registered using the minimum eigenvalue criteria (R package 'fda'). **Lower Left** Boys velocity functions registered by F-R (R package 'fdasrvf'). **Lower Right** Boys velocity functions registered by the GP model.

respect, the underlying registration principle driving our model is similar to that proposed by Ramsay and Li [34]. Here we will compare our registration results to those using Ramsey's method as well as the registration procedure proposed by Srivistava, et.al. [43]. Srivistava et. al. propose a geometric framework for functional data registration using the Fisher-Rao Riemannian metric, Rao [37]. In this paper we will refer to Ramsey and Li's registration procedure as ME (minimum eigenvalue) and Srivistava's procedure as F-R (Fisher-Rao), and the model proposed here as GP (Gaussian Processes). In the paper by Srivistava, et. al., they provide several comparisons of registration under the F-R framework to the registration methods proposed by Gervini and Gasser [13], James [17], Liu and Müller [26], and Tang and Müller [44]. In all cases, F-R appears to provide the most complete registration of the given set of functions. In light of this illustration, we will consider their method as the current frontrunner in registration procedures and use it as the standard for our comparisons.

In Figures 3.1, 3.2, and 3.3 are the three datasets that are used for this analysis. Each figure contains the original unregistered data along with plots of the functions registered using the three proposed methods. For all three registration methods, a range of parameter values were explored for optimal registration.

We have chosen to use the Sobolev Least Squares (*sls*) criterion to compare the three sets of registered functions. The Sobolev Least Squares criterion compares the total cross-sectional variance of the first derivatives of the registered functions to that of the original functions. Explicitly,

$$sls = \frac{\sum_{i=1}^{N} \int (X_i'(h_i(t)) - \frac{1}{N}\sum_{j=1}^{N} X_j'(h_j(t)))^2 dt}{\sum_{i=1}^{N} \int (X_i'(t) - \frac{1}{N}\sum_{j=1}^{N} X_j'(t))^2 dt} \tag{3.9}$$

In Srivistava, et.al. [43] *sls* is seen as the best measure of alignment in comparison to two other criterion, a least squares criterion and a pairwise correlation criterion. Lower values of *sls* correspond to better function alignment.

First Simulated Data Set Figure 3.1 contains the functions of the first simulated data set. These data consist of 18 shifted and scaled Gaussian probability density functions. For this simple registration there is not much difference between the three registration procedures.

Second Simulated Data Set Figure 3.2 contains the functions of the second simulated data set. These data consist of 20 unregistered scaled mixtures of three Gaussian probability density functions. Again all three registration procedures result in similar alignments. However, the GP method does a better job of recovering the original shape of the functions and results in the lowest *sls*. Note: The ME registered functions are based on 5 complete runs of the ME algorithm where in each run the previous runs results were used as the 'unregistered' functions.

Berkeley Boys Growth Velocity Data Figure 3.3 contains 39 velocity of growth functions for boys from the Berkeley Growth Study, Tuddenham and Snyder [46]. For this analysis, the original data are slightly changed to eliminate some erratic behavior at the beginning of each function. Here, again, GP and F-R yield similar registration results. However, the GP algorithm results in the lowest *sls*. ME registers the most significant peak in growth velocity but does not align lesser features as well as GP. Note: The ME registered functions are based on 2 complete runs of the ME algorithm where in each run the previous runs results were used as the 'unregistered' functions. Running this algorithm more than twice resulted in function distortion due to over-warping and a larger

sls.

While the GP and F-R methods result in a similar alignment of functions, these results are achieved in very different environments that are specialized to satisfy specific inferential preferences. The F-R registration method is convenient (using R package 'fdasrvf') and provides fast high-quality estimates. On the other hand, while providing comparable registration results, our method expands inferential capability by providing 1) variability estimates for all unknown parameters and 2) a probability framework in which future partially observed unregistered functions are considered. In contrast to traditional functional prediction methods, our model not only provides an estimate of the complete unregistered function, but also estimates the complete warping function and the complete registered function. Details of the prediction model are found in Section 3.4.

3.3.2 Comparison to MCMC Results

To establish the utility of the adapted variational Bayes algorithm, here we compare the estimates of registered functions using adapted variational Bayes versus those obtained through MCMC sampling. For this exposition, the two simulated data sets and the Boys Growth Velocity data set described in Section 3.3.1 are used to look at the discrepancies between the estimated registered functions from MCMC sampling versus those determined by the AVB algorithm.

The squared L^2 norm of the difference between the AVB and MCMC estimate of a registered function is used to quantify the differences between these estimates. Figure 3.4 illustrates for both simulated data sets how closely the

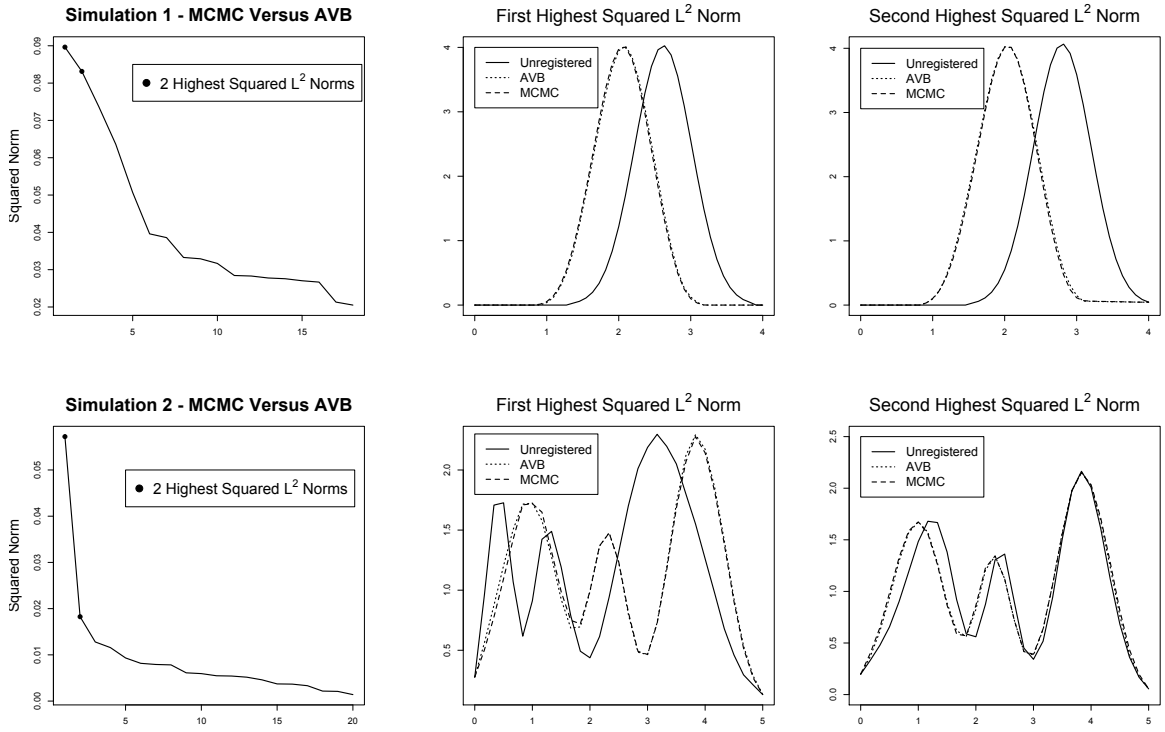

Figure 3.4: Simulations 1 and 2 - Differences Between MCMC and AVB Estimates. **Top and Lower Left** Plot of the squared L^2 norm of the difference between the MCMC and AVB estimates for each observation in decreasing order of magnitude for simulated data sets 1 and 2 respectively. **Top Center and Left** The original unregistered function plotted with the MCMC and AVB estimates of the registered functions for the observations from simulated data set 1 with the two largest discrepancies between the MCMC and AVB estimates. **Lower Center and Left** The original unregistered function plotted with the MCMC and AVB estimates of the registered functions for the observations from simulated data set 2 with the two largest discrepancies between the MCMC and AVB estimates.

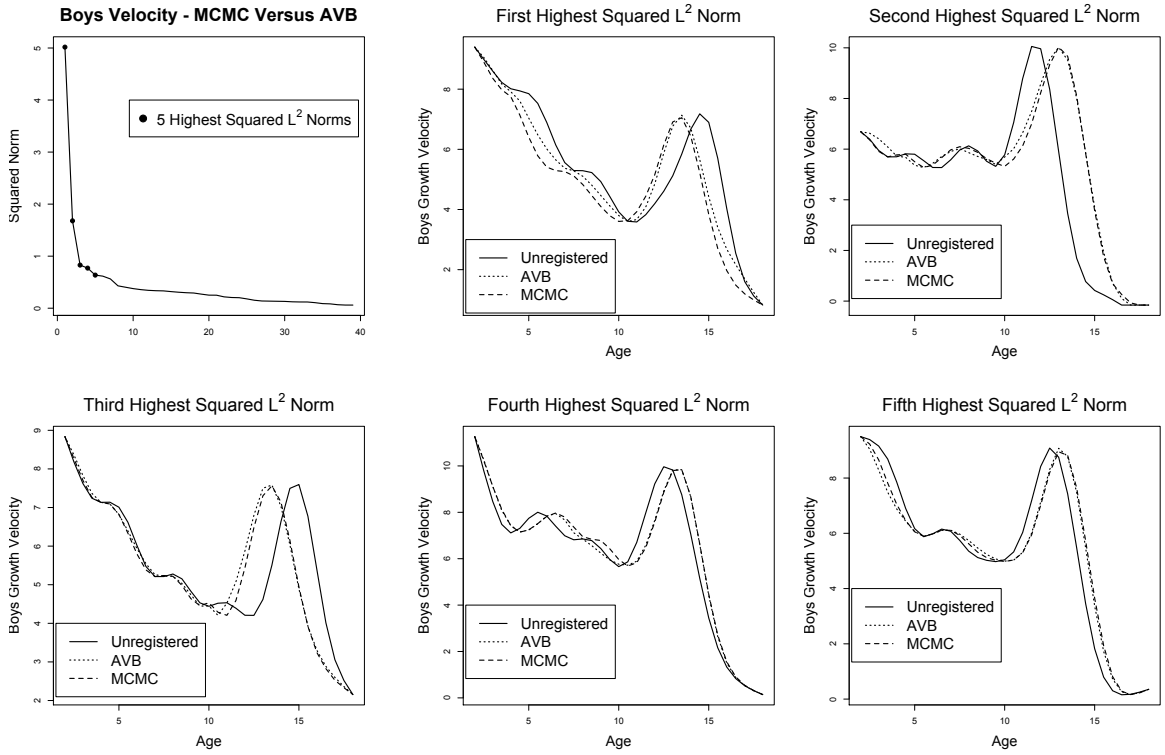

Figure 3.5: Registered Boys Growth Velocity - Differences Between MCMC and AVB Estimates. **Top Left** Plot of the squared L^2 norm of the difference between the MCMC and AVB estimates for each observation in decreasing order of magnitude . **Top Center and Left** The original unregistered function plotted with the MCMC and AVB estimates of the registered functions for the observations with the first two largest discrepancies between the MCMC and AVB estimates. **Lower** Plots of the next three observations with the highest squared L^2 norms of the difference between the MCMC and AVB estimates. The squared L^2 norm associated with the lower right plot is about .64. As can be seen in this illustration, at this level there are only small differences between the MCMC and AVB estimates.

AVB estimates follow the MCMC estimates. Even the largest squared L^2 norms of the differences between these two estimates correspond to minor changes in the estimates. These simulations represent rather ideal conditions for registration where there is almost no variation in the registered functions beyond a scaling and vertical shift of the target function. Consequently, as we might expect, the MCMC and AVB estimation procedures are primarily in agreement. Figure 3.5 is a more realistic look at the differences between the MCMC and AVB registration results for data that has significant variation in the registered functions beyond a scaling and vertical shift of the target function. However, even here we see the AVB algorithm performs well. Of the 39 observations, in only 2 or 3 are there notable discrepancies between the AVB and MCMC estimated registered functions.

3.4 Variational Approximation for Functional Prediction

3.4.1 Functional Prediction with Bootstrapped Credible Intervals

The probabilistic framework of our registration model provides a natural structure in which we can consider new observations. Functional prediction has been considered by Ferraty and Vieu [12]. Here we extend current methods by taking into account the phase variability of a partially observed function.

We will make the following assumptions:

1. We have a sample of approximated unregistered functions, $\mathbf{X}_i$ =

$(X_i(t_1), \ldots, X_i(t_p))'$, $i = 1, \ldots, N$.

2. $\mathbf{X}_i = (X_i(t_1), \ldots, X_i(t_p))'$, $i = 1, \ldots, N$ are registered using the registration method outlined in Section 3.1 via a MCMC sampler or adapted variational Bayes.

3. From (2) we have obtained estimates for the target function, $f(t)$, the registered functions, $X_i(h_i(t))$, $i = 1, \ldots, N$, the warping functions, $h_i(w_i(t))$, $i = 1, \ldots, N$, σ^2_{z0}, and σ^2_{z1}.

4. A new function, $X_{N+1}(t)$ has been observed at the time points $(t_1, \ldots, t_r)'$, $r < p$.

5. $(X(h(t_1)), \ldots X(h(t_p)))' \approx N_p(\hat{\boldsymbol{\mu}}_{\mathbf{X(h)}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{X(h)}})$, the distribution of the registered functions can be approximated by a multivariate normal distribution using the sample mean, $\hat{\boldsymbol{\mu}}_{\mathbf{X(h)}}$, and sample covariance matrix, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X(h)}}$, of the estimated registered functions obtained in (2).

6. $(w(t_1), \ldots w(t_{p-1}))' \approx N_{p-1}(\hat{\boldsymbol{\mu}}_{\mathbf{w}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{w}})$, the distribution of the base functions can be approximated by a multivariate normal distribution using the sample mean, $\hat{\boldsymbol{\mu}}_{\mathbf{w}}$, and sample covariance matrix, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{w}}$, of the estimated base functions obtained in (2).

Under these assumptions, we will proceed as follows:

1. Register the partially observed function, $\mathbf{X^P}_{N+1} = (X_{N+1}(t_1), \ldots, X_{N+1}(t_r))'$ to the estimated target function, $\hat{f}(t)$, truncated to an appropriate registration time, t_f, $f \in \{1, \ldots, p\}$, so that $h_{N+1}(t_f) = t_r$.

2. Using the distributions from assumptions (5) and (6) above, the estimate of the partial registered function, $\mathbf{X^P}_{N+1}(\hat{\mathbf{h}}_{N+1}) = (X^P_{N+1}(\hat{h}_{N+1}(t_1)), \ldots, X^P_{N+1}(\hat{h}_{N+1}(t_f)))'$ and the estimate of the partial base function, $\widehat{\mathbf{w^P}}_{N+1} = (\widehat{w^P}_{N+1}(t_1), \ldots \widehat{w^P}_{N+1}(t_{f-1}))'$,

estimate the registered and base functions to time t_p and t_{p-1} respectively using the conditional properties of the multivariate normal distribution. Accordingly, denoting future registered observations and future warping function values , $\mathbf{X^F}_{N+1}(\mathbf{h}_{N+1})$ and $\mathbf{w^F}_{N+1}$, respectively, the estimates of these future values are
$\widehat{\mathbf{X^F}}_{N+1}(\hat{\mathbf{h}}_{N+1}) = E(\mathbf{X^F(h)}|\mathbf{X^P}_{N+1}(\hat{\mathbf{h}}_{N+1}), \hat{\boldsymbol{\mu}}_{\mathbf{X(h)}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{X(h)}})$ and
$\widehat{\mathbf{w^F}}_{N+1} = E(\mathbf{w^F}|\widehat{\mathbf{w^P}}_{N+1}, \hat{\boldsymbol{\mu}}_{\mathbf{w}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{w}})$.

3. Estimate the complete unregistered function, $X_{N+1}(t)$, using the inverse of the estimated warping function and the estimated registered function.

An additional random element in the prediction model is the last registered time of the truncated target function, t_f, used to register the partial observation. To obtain the best possible registration of the partial observation, a range of final registration times are considered over a finer domain. The efficiency of the adapted variational Bayes algorithm makes it possible to consider several possible partial registrations as follows:

1. For each of the time points t_j, $j \in \{m, \ldots, (m+k-1)\}$, $t_{m+k-1} < t_p$, the partially observed function, $X^P_{N+1}(t)$, is registered to the estimated target function, $\hat{f}(t)$, truncated to time, t_j, so that $\hat{h}_{N+1(j)}(t_j) = t_r$, where $\hat{h}_{N+1(j)}(t)$ is the estimated warping function determined by registering $X^P_{N+1}(t)$ to the proposed final registration time t_j. Note, the first and last times considered in this interval are chosen by plotting the partial unregistered function and the target function together and determining a generous interval that contains the appropriate final registration time. This interval is subsequently made finer to allow this time to fall between two of the original time points.

2. Calculate $d_{t_j} = \|\mathbf{X^P}_{N+1} - (\hat{z}_{0(j)}\mathbf{1} + \hat{z}_{1(j)}\mathbf{f^U}_{(j)})\|_2$ for each t_j, $j \in \{m, \ldots, m+k-1\}$ where

   $\mathbf{f^U}_{(j)} = (\hat{f}(t_1), \hat{f}(\hat{h}^{-1}_{N+1(j)}(t_2)), \ldots, \hat{f}(\hat{h}^{-1}_{N+1(j)}(t_r) = t_j))'$

3. $t_f = \underset{t_j, j\in\{m,\ldots,m+k-1\}}{\arg\min} d_{t_j}$

This algorithm determines the final registered time that results in the minimum L^2 norm between the partially recorded unregistered function and the target function evaluated at the inverse of the warping function estimated using that final time. Note, for all j, $\mathbf{f^U}_{(j)}$ shares the same domain as the partially recorded unregistered function, $\mathbf{X^P}_{N+1}$.

The efficiency of adapted variational Bayes for prediction also makes it possible to characterize variability in the estimates of the complete registered function, unregistered function, and base function via bootstrapping. For M bootstrapped samples of the predicted registered functions, warping functions, and unregistered functions, for $m = 1, \ldots, M$

1. Draw a new sample of registered functions from $N_p(\hat{\boldsymbol{\mu}}_{\mathbf{X(h)}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{X(h)}})$
2. Draw a new sample of base functions from $N_{p-1}(\hat{\boldsymbol{\mu}}_{\mathbf{w}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{w}})$
3. Use these samples for prediction steps (1) - (3) above to determine bootstrapped estimates of the complete, registered, unregistered, and base functions.

3.4.2 Functional Prediction - El-Niño Data

The El-Niño data consist of weekly readings of sea surface temperature with the first observation in June of 1950. Complete data can be found at NOAA's Cli-

mate Prediction Center website (http://www.cpc.ncep.noaa.gov/data/indices/). The data that we are using for this analysis are found through Professor Frederic Ferraty's (Mathematics; University of Toulouse, France) website (http://www.math.univ-toulouse.fr/ ferraty/SOFTWARES/NPFDA/npfda-datasets.html). These data are a subset of the original data with monthly sea surface temperature records from June of 1950 to May of 2004. For this analysis, the bi-monthly observations are added to the data to prevent significant changes to the shape of a given function due to interpolation error. Also, light smoothing is applied to all functions.

One of the climatological motives for recording sea surface temperatures is to monitor the El-Niño phenomenon. El-Niño conditions and episodes are characterized by a prolonged increase of at least .5 °C in sea surface temperatures from the average sea surface temperature. El-Niño conditions affect weather patterns and water conditions especially along the coastlines. Common weather effects are flooding, abnormally dry or wet weather, and changes in tropical storm paths. Changing water conditions caused by the El-Niño phenomenon diminish large fish populations which impacts local fishing and international markets, Wikipedia Contributers [49]. The goal of our study is to predict how high temperatures will stay in the remaining part of the year in conjunction with how long temperatures will drop before they rise again based on the first seven months of temperature recordings from the lowest temperature recording in the previous year.

For this purpose, the data are restructured to define a"year" as the period of time between the lowest temperatures in consecutive calendar years. For example, the first year in our data set ranged from September 1950 to September

1951. Note, these "years" will not all be 12 months in length, and our final data had "years" that ranged from 11 to 14 months. After splitting the data into years based on this definition, the observations were split into 3 groups based on the previous year's lowest temperature. The first temperature group consists of years where the previous year's lowest temperature was less than or equal to 19.5 °C. The second temperature group consists of years where the previous year's lowest temperature was greater than 19.5 °C and less than 21 °C. The final group consists of years where the previous year's lowest temperature was greater than or equal to 21 °C. The sea surface temperature profiles within each group are more similar to each other than those between groups as years that start particularly cold tend to be less cold in the next year and years with a milder start tend to be colder in the following year. For our analysis, we will concentrate on the second group of functions characterized by a previous year's lowest temperature greater than 19.5 °C and less than 21 °C. This group contains 29 functions. The first 28 functions will be used to predict the remaining portion of the 29th function based on the first 7 months of sea surface temperature observed in that year.

For the purpose of registration, all functions need to be recorded over the same interval of time. As mentioned above, in this particular case our data is recorded over a time periods that range from 11 to 14 months. An easy remedy to this situation is to perform a simple initial warping to each function that rescales every observation to an 11 month time frame. In our final analysis, this initial warping is accounted for when determining the final base functions used for the prediction algorithm.

The original unregistered functions and the functions registered using the

GP model described in Section 3.1 are plotted in Figure 3.6. For this data set, to register significant features in the sample while retaining function variation beyond a scaling and vertical shift of the target function, individual warping parameters, γ_{w_i}, $i = 1, \ldots, 28$ were utilized instead of γ_w in (3.6). Significant differences in the amplitude variation in the original functions that is unassociated with temporal variation prevented the use of a global parameter. However, only 3 unique warping parameters in total were necessary.

Using the empirical mean of the 28 original registered functions as the target function, the first 7 months of sea surface temperature records from observation 29 are registered to a piece of the target function where the final registered time is allowed to vary from 6.5 to 7.5 months. Between these months, a finer time interval corresponding to weekly records is used to allow for additional flexibility in determining the final registered time. The partially recorded function is plotted with the target function in the lower right panel of Figure 3.6. The grey shaded area includes the time points considered for the final time of the partial registration. After the optimal registration of the partially recorded observation is determined, estimates of the entire registered function, warping function, and unregistered function are determined using the model outlined in Section 3.4.1.

One-hundred bootstrapped samples were used to estimate the variability in the predictions of all three estimated functions. Figure 3.7 plots the initial estimates with the 95% bootstrapped confidence intervals. In addition, the plot of the estimated unregistered function also includes the true value of this function.

The primary advantage of registering the partially recorded observation before estimating future values is that we can capture variation in amplitude and timing separately. In Figure 3.7, the first plot captures the variability in the fu-

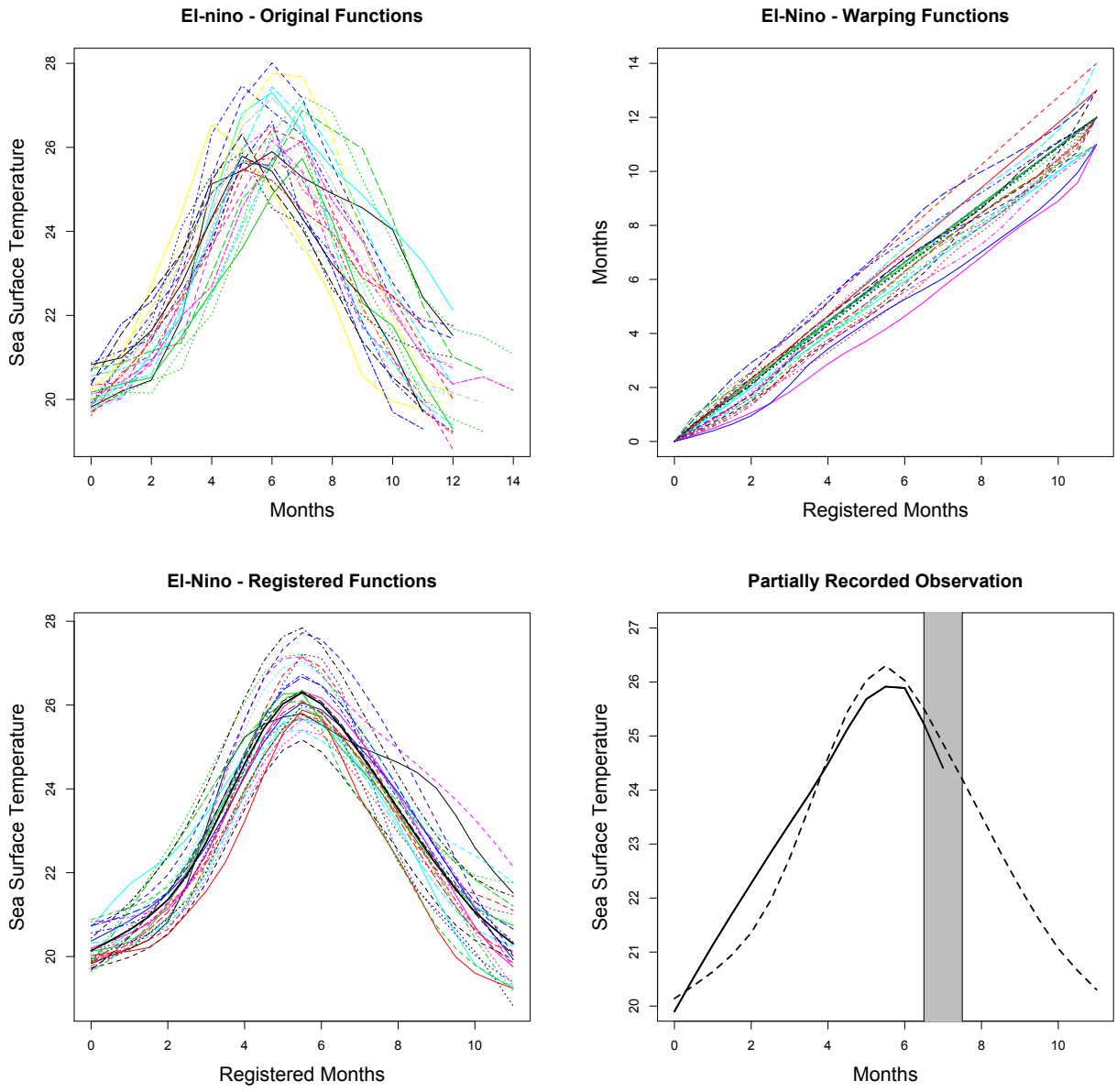

Figure 3.6: El-niño Data. **Top Left** Original 28 profiles of sea surface temperature. **Top Right** Estimated warping functions. As can be seen here, the time period of the original data ranged from 11 to 14 months. **Lower Left** Estimated registered temperature profiles. **Lower Right** The solid line is observation 29 recorded for 7 months. The dashed line is the estimated target function. The grey shaded area spans the 5 time points that are considered for the final time of the partial registration.

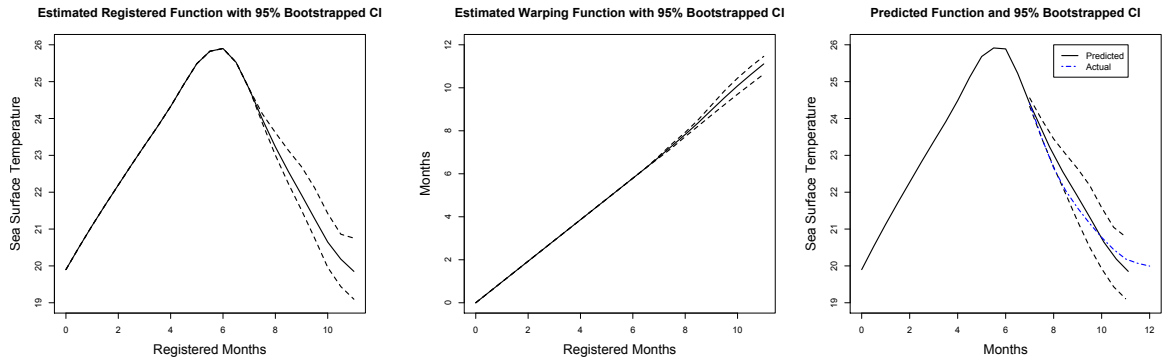

Figure 3.7: Estimates and Bootstrapped Confidence Intervals. **Left** Estimated registered function with 95% bootstrapped confidence interval. **Center** Estimated warping function with 95% bootstrapped confidence interval. **Right** Estimated unregistered function with 95% bootstrapped confidence interval. The dashed and dotted line is the true unregistered function.

ture level of sea surface temperature (amplitude variation), and answers the question,"How high can we expect sea surface temperatures to stay?". The second plot captures the variability in the timing of future observations (temporal variation) which addresses the question of, "When can we expect sea level temperatures to begin rising again?". The confidence interval for the unregistered function seen in the last frame of Figure 3.7, combines both amplitude and temporal variation to estimate the future trajectory of sea surface temperature for this year. In this illustration it can be seen that the main difference in the estimated and actual temperature profiles lies in the timing of the lowest observation. However, for this observation, the sea surface temperature at 12 months is not much different than the sea surface temperature at 11.5 months. The predicted timing of the lowest temperature was 11.1 months.

One of the most notable features of this analysis is that there is little uncertainty in the registration of the first 7 months of sea surface temperatures. The

most prominent feature in the data is the peak temperature that occurs anywhere from 4 to 8 months in the original data. In our partially recorded observation, as seen in Figure 3.6, the peak of the target function and the partially recorded observation are already closely aligned. Additionally, this observation happens to be similar in shape to the target function. The combination of these features resulted in only a minimal amount of variation in the estimated registered and warping functions in the first 7 months. However, we note here, this phenomenon is an artifact of these particular data, and in other analyses more variation in the registered timing of the partially recorded observation would be expected.

The El-niño data set provides a challenging registration problem. The registered functions vary significantly in directions beyond the target function. Choosing curve specific registration parameters enabled features common to all functions to be registered while retaining prominent features in each individual curve. This is just one example of the difficulties that can arise in registering functional data and in turn how these challenges can be addressed to analyze data that does not fit the "ideal" registration problem.

3.5 Functional Data Regularization and Registration

3.5.1 Combining Registration and Smoothing

If instead of the function itself, noisy observations of each unregistered function, $X_i(t)$ are observed over a finite number of time points, $\mathbf{t} = (t_1, \ldots, t_p)'$, we will additionally assume that the observations, $Y_i(t_j)$, $j = 1, \ldots, p$ are iid, $N(X_i(t_j), \sigma_Y^2)$.

Incorporating registration and smoothing into a single model has also been considered recently by Rakêt et. al. [31]. In their paper, each registered noisy function, $Y_i(h_i(t))$ at time point t_j, $j = 1, \ldots, p$ is composed as follows:

$$Y_i(h_i(t_j)) \quad = \quad f(t_j) + r_i(t_j) + \epsilon_i(t_j)$$

where $f(t)$ is similar to our target function, $r_i(t)$ is a function-specific random effect that accounts for variation in individual noiseless functions beyond the target function, and $\epsilon_i(t_j)$ are iid Gaussian noise.

The advantage of our model is that incorporating individual random effects is unnecessary. Noting that the *observations* are noisy, not the registered functions; smoothing in our model is applied to the observations, not to the functions after registration. Under these conditions, variability in the estimated unregistered, smoothed functions, $X_i(t)$, can be looked at separately from variability in the estimated registered functions, $X_i(h_i(t))$. Section 3.5.3 provides an example of how treating smoothing as a pre-processing step underestimates variability in the posterior distributions of the registered functions.

In the presence of noisy observations, the following distributions are either altered or added to the registration model presented in Section 3.1.

$$\begin{aligned}
Y_i(t_j) \mid X_i(t_j) &\sim N(X_i(t_j), \sigma_Y^2), \quad i = 1, \ldots, N \quad j = 1, \ldots, p && (3.10)\\
\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}, \eta_X, \lambda_X &\sim N_p(z_{0i}\mathbf{1} + z_{1i}\mathbf{f}, \gamma_R^{-1}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_X) \quad i = 1, \ldots, N && (3.11)\\
\boldsymbol{\Sigma}_X &= \eta_X^{-1}\mathbf{P}_1 + \lambda_X^{-1}\mathbf{P}_2\\
\eta_X &\sim G(c, d)\\
\lambda_X &\sim G(c, d)\\
\sigma_Y^2 &\sim IG(a, b)
\end{aligned}$$

The most significant change to the model is that we now include a smoothing penalty in the covariance specification for the registered functions. Here specifying $\mathbf{P}_2$ in the prior distribution for the registered functions establishes regularization in these functions. The associated smoothing parameter is λ_X. As mentioned previously, $\mathbf{P}_1$ is required to define $\mathbf{\Sigma}_X$ as a proper covariance matrix. More details on these matrices can be found in Section 2.2.1. As can be seen above, η_X and λ_X can be selected through the inference procedure and are considered as additional unknown parameters.

In the prior specifications of this model, equation (3.11) incorporates penalties for both smoothing and registration within the prior for the registered functions. The full conditional distribution for each approximated registered function, $\mathbf{X}_i(\mathbf{h}_i)$, when data are noisily observed is the joint full-conditional of the unregistered function and the warping function.

$$\begin{aligned} f(\mathbf{X}_i(\mathbf{h}_i) \mid rest) &= f(\mathbf{w}_i, \mathbf{X}_i \mid rest) \\ &= f(\mathbf{w}_i \mid \mathbf{X}_i, rest) f(\mathbf{X}_i \mid rest) \end{aligned}$$

Instead of drawing from this joint full-conditional directly, we will proceed by first drawing from $f(\mathbf{X}_i \mid rest)$ and then given $\mathbf{X}_i$, draw from $f(\mathbf{w}_i \mid \mathbf{X}_i, rest)$.

These full conditional distributions are determined in the standard way recognizing that the prior distribution for a registered function can be factored into two components. One component penalizes lack of registration given the approximated unregistered function, $\mathbf{X}_i$; the other component penalizes roughness in the registered function which implicitly penalizes roughness in the unregistered function. The roughness penalty is independent of the warping function and therefore also of $\mathbf{w}_i$. Specifically, the prior distribution (3.11) for each $\mathbf{X}_i(\mathbf{h}_i)$,

$i, \ldots, N$, is such that

$$f(\mathbf{X}_i(\mathbf{h}_i) \mid \mathbf{X}_i, \mathbf{w}_i, z_{0i}, z_{1i}, \mathbf{f}, \eta_X, \lambda_X) \propto$$

$$exp\left[-\frac{1}{2}\left((\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}))'\gamma_R\mathbf{\Sigma}^{-1}(\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}))\right)\right] * \tag{3.12}$$

$$exp\left[-\frac{1}{2}\left((\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}))'\mathbf{\Sigma}_X^{-1}(\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}))\right)\right] \tag{3.13}$$

Accordingly, the components of the joint distribution of the data and all unknown parameters that are dependent on $\mathbf{w}_i$ are expressions (3.6) and (3.12), and the resulting full conditional distribution for the approximated functions $\mathbf{w}_i$ is such that

$$f(\mathbf{w}_i \mid rest) \propto$$

$$exp\left[-\frac{1}{2}\left((\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}))'\gamma_R\mathbf{\Sigma}^{-1}(\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}))\right)\right] *$$
$$exp\left[-\frac{1}{2}\left(\mathbf{w}_i'\gamma_w\mathbf{\Sigma}^{-1}\mathbf{w}_i\right)\right]$$

This full conditional does not have a known distributional form and can be sampled from via a Metropolis step in a MCMC sampler.

The components of the joint distribution of the data and all unknown parameters that are dependent on $\mathbf{X}_i$ is the data distribution (3.10) and expression (3.13). The resulting full conditional distribution is such that

$$f(\mathbf{X}_i \mid rest) \propto$$

$$exp\left[-\frac{1}{2\sigma_Y^2}(\mathbf{Y}_i - \mathbf{X}_i)'(\mathbf{Y}_i - \mathbf{X}_i)\right] *$$
$$exp\left[-\frac{1}{2}\left((\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}))'\mathbf{\Sigma}_X^{-1}(\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}))\right)\right]$$

This full conditional distribution also is not of a known distributional form and can be sampled from using a Metropolis step. However, as significant features of the unregistered function, $\mathbf{X}_i$, should be unchanged by the registration. Smoothness in the registered function, $\mathbf{X}_i(\mathbf{h}_i)$, implies the same level of smoothness in the unregistered function $\mathbf{X}_i$. For ease of sampling, we will re-write (3.13) in terms of the unregistered function, $\mathbf{X}_i$ so that

$$
\begin{aligned}
f(\mathbf{X}_i \mid rest) \propto & \\
& exp\left[-\frac{1}{2\sigma_Y^2}(\mathbf{Y}_i - \mathbf{X}_i)'(\mathbf{Y}_i - \mathbf{X}_i)\right] * \\
& exp\left[-\frac{1}{2}\left((\mathbf{X}_i - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}(\mathbf{h}_i^{-1})))'\mathbf{\Sigma}_X^{-1}(\mathbf{X}_i - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}(\mathbf{h}_i^{-1})))\right)\right] \qquad (3.14)
\end{aligned}
$$

which results in a multivariate normal full conditional distribution for $\mathbf{X}_i$.

When noisy observations, $\mathbf{Y}_i$, $i = 1 \ldots N$ are recorded, the approximation we make in (3.14), while preserving conjugacy, prevents exact variational Bayes updates to be performed on the approximate posterior distributions for the following parameters: $\mathbf{X}_i$, $i = 1 \ldots N$, σ_Y^2, η_X, and λ_X. Hence, the adapted variational Bayes procedure proposed here requires special handling under this data assumption.

3.5.2 Adapted Variational Bayes For Noisy Functional Data

When the functional data are recorded with noise, the adapted variational Bayes algorithm requires further adjustments to perform an approximate inference procedure. With the necessary adjustments, the convergence properties of the adapted variational Bayes algorithm no longer hold. However, we have found in practice that the adjusted algorithm still results in useful estimates for initial-

izing a MCMC sampler.

Here we look at why the approximate posterior distributions for $\mathbf{X}_i$, $i = 1, \ldots, N$, η_X, and λ_X cannot be updated properly using the adapted variational Bayes algorithm. In the m^{th} iteration, the following update should be made to $\log q(\mathbf{X}_i)$, for $i = 1 \ldots N$:

$$log\ [q^{(m)}(\mathbf{X}_i)] \propto E_{(\theta_{-\mathbf{X}_i})}[log\ f(\mathbf{X}_i \mid rest)]$$

where

$$\begin{aligned} E_{(\theta_{-\mathbf{X}_i})}[log\ f(\mathbf{X}_i \mid rest)] &\propto E_{(\theta_{-\mathbf{X}_i})}\Big[-\frac{1}{2}[(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}_i|rest})' \boldsymbol{\Sigma}^{-1}_{\mathbf{X}_i|rest}(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}_i|rest})]\Big] \\ \boldsymbol{\Sigma}_{\mathbf{X}_i|rest} &= (\frac{1}{\sigma^2_Y}\mathbf{I}_p + \eta_X \mathbf{P}_1^- + \lambda_X \mathbf{P}_2^-)^{-1} \\ \boldsymbol{\mu}_{\mathbf{X}_i|rest} &= \boldsymbol{\Sigma}_{\mathbf{X}_i|rest}[\frac{1}{\sigma^2_Y}\mathbf{Y}_i + (\eta_X \mathbf{P}_1^- + \lambda_X \mathbf{P}_2^-)(z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}(\mathbf{h}_i^{-1}))] \end{aligned}$$

Taking the expectation over the q distributions for all other parameters except for the base functions results to the following updated parameters of $q(\mathbf{X}_i) = N_p(\boldsymbol{\mu}_{q(\mathbf{X}_i)}, \boldsymbol{\Sigma}_{q(\mathbf{X}_i)})$

$$\begin{aligned} \boldsymbol{\Sigma}^{(m)}_{q(\mathbf{X}_i)} &= (\mu_{q(\frac{1}{\sigma^2_Y})}\mathbf{I}_p + \mu_{q(\eta_X)}\mathbf{P}_1^- + \mu_{q(\lambda_X)}\mathbf{P}_2^-)^{-1} \\ \boldsymbol{\mu}^{(m)}_{q(\mathbf{X}_i)} &= \boldsymbol{\Sigma}^{(m)}_{q(\mathbf{X}_i)}[\mu_{q(\frac{1}{\sigma^2_Y})}\mathbf{Y}_i + (\mu_{q(\eta_X)}\mathbf{P}_1^- + \mu_{q(\lambda_X)}\mathbf{P}_2^-)(\mu_{q(z_{0i})}\mathbf{1}_p + \mu_{q(z_{1i})}E_{(\theta_{-\mathbf{X}_i})}[\mathbf{f}(\mathbf{h}_i^{-1})])] \end{aligned} \tag{3.15}$$

In (3.15), the expectation of $\mathbf{f}(\mathbf{h}_i^{-1})$ is unknown. So, the first approximation we will make is that $E_{(\theta_{-\mathbf{X}_i})}[\mathbf{f}(\mathbf{h}_i^{-1})] \approx \boldsymbol{\mu}_{q(\mathbf{f})}(\mathbf{h}_i^{-1})$.

Similarly, to update $\log q(\eta_X)$:

$$log\ [q^{(m)}(\eta_X)] \propto E_{(\theta_{-\eta_X})}[log\ f(\eta_X \mid rest)]$$

where

$$
\begin{aligned}
E_{(\theta_{-\eta_X})}[log\ f(\eta_X \mid rest)] &\propto E_{(\theta_{-\eta_X})}[c_{\eta_X|rest} log\ \eta_X - d_{\eta_X|rest}\eta_X] \\
c_{\eta_X|rest} &= N + c \\
d_{\eta_X|rest} &= d + \frac{1}{2}\sum_{i=1}^{N} tr[(\mathbf{X}_i - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}(\mathbf{h}_i^{-1})))(\mathbf{X}_i - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}(\mathbf{h}_i^{-1})))'\mathbf{P}_1^{-}]
\end{aligned}
$$

Taking the expectation over the q distributions for all other parameters except for the base functions results to the following updated parameters of $q(\eta_X) = G(c_{q(\eta_X)}, d_{q(\eta_X)})$,

$$
\begin{aligned}
c_{q(\eta_X)}^{(m)} &= N + c \\
d_{q(\eta_X)}^{(m)} &= d + \frac{1}{2} tr\Big[\Big(\sum_{i=1}^{N}\Big(\boldsymbol{\Sigma}_{q(\mathbf{X}_i)} + \boldsymbol{\mu}_{q(\mathbf{X}_i)}\boldsymbol{\mu}'_{q(\mathbf{X}_i)} - 2\boldsymbol{\mu}_{q(\mathbf{X}_i)}(\mu_{q(z_{0i})}\mathbf{1}_p + \mu_{q(z_{1i})}E_{(\theta_{-\eta_X})}[\mathbf{f}(\mathbf{h}_i^{-1})])' \\
&+ 2\mu_{q(z_{0i})}\mu_{q(z_{1i})}\mathbf{1}_p E_{(\theta_{-\eta_X})}[\mathbf{f}(\mathbf{h}_i^{-1})]' + (\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})E_{(\theta_{-\eta_X})}[\mathbf{f}(\mathbf{h}_i^{-1})\mathbf{f}(\mathbf{h}_i^{-1})']\Big) \\
&+ \Big(2\sum_{i=1}^{N-1}(\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})}) + \sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\mu_{q(z_{0i})}\mu_{q(z_{0j})}\mathbb{1}\{j \neq i\}\Big)\mathbf{1}_p\mathbf{1}'_p\Big)\mathbf{P}_1^{-}\Big]
\end{aligned}
$$

In the expression for $d_{q(\eta_X)}^{(m)}$, $E_{(\theta_{-\eta_X})}[\mathbf{f}(\mathbf{h}_i^{-1})]$ and $E_{(\theta_{-\eta_X})}[\mathbf{f}(\mathbf{h}_i^{-1})\mathbf{f}(\mathbf{h}_i^{-1})']$ are unknown. Thus, we will make the following approximations

$$E_{(\theta_{-\eta_X})}[\mathbf{f}(\mathbf{h}_i^{-1})] \approx \boldsymbol{\mu}_{q(\mathbf{f})}(\mathbf{h}_i^{-1}) \text{ and } E_{(\theta_{-\eta_X})}[\mathbf{f}(\mathbf{h}_i^{-1})\mathbf{f}(\mathbf{h}_i^{-1})'] \approx \boldsymbol{\Sigma}_{q(\mathbf{X}_i)}/N + \boldsymbol{\mu}_{q(\mathbf{f})}(\mathbf{h}_i^{-1})\boldsymbol{\mu}_{q(\mathbf{f})}(\mathbf{h}_i^{-1})'$$

Note, $\boldsymbol{\Sigma}_{q(\mathbf{X}_i)}$ does not depend on i.

The variational Bayes algorithm update for λ_X is similar to that of η_X and requires the same approximations.

$$log\ [q^{(m)}(\lambda_X)] \propto E_{(\theta_{-\lambda_X})}[log\ f(\lambda_X \mid rest)]$$

where

$$
\begin{aligned}
E_{(\theta_{-\lambda_X})}[log\ f(\lambda_X \mid rest)] &\propto E_{(\theta_{-\lambda_X})}[c_{\lambda_X|rest} log\ \lambda_X - d_{\lambda_X|rest}\lambda_X] \\
c_{\lambda_X|rest} &= N\Big(\frac{p-2}{2}\Big) + c \\
d_{\eta_X|rest} &= d + \frac{1}{2}\sum_{i=1}^{N} tr[(\mathbf{X}_i - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}(\mathbf{h}_i^{-1})))(\mathbf{X}_i - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}(\mathbf{h}_i^{-1})))'\mathbf{P}_2^-]
\end{aligned}
$$

Taking the expectation over the q distributions for all other parameters except for the base functions results to the following updated parameters of $q(\lambda_X) = G(c_{q(\lambda_X)}, d_{q(\lambda_X)})$,

$$
\begin{aligned}
c_{q(\lambda_X)}^{(m)} &= N\Big(\frac{p-2}{2}\Big) + c \\
d_{q(\lambda_X)}^{(m)} &= d + \frac{1}{2} tr\Big[\Big(\sum_{i=1}^{N}\Big(\boldsymbol{\Sigma}_{q(\mathbf{X}_i)} + \boldsymbol{\mu}_{q(\mathbf{X}_i)}\boldsymbol{\mu}'_{q(\mathbf{X}_i)} - 2\boldsymbol{\mu}_{q(\mathbf{X}_i)}(\mu_{q(z_{0i})}\mathbf{1}_p + \mu_{q(z_{1i})}E_{(\theta_{-\lambda_X})}[\mathbf{f}(\mathbf{h}_i^{-1})])' \\
&+ 2\mu_{q(z_{0i})}\mu_{q(z_{1i})}\mathbf{1}_p E_{(\theta_{-\lambda_X})}[\mathbf{f}(\mathbf{h}_i^{-1})]' + (\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})E_{(\theta_{-\lambda_X})}[\mathbf{f}(\mathbf{h}_i^{-1})\mathbf{f}(\mathbf{h}_i^{-1})']\Big) \\
&+ \Big(2\sum_{i=1}^{N-1}(\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})}) + \sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\mu_{q(z_{0i})}\mu_{q(z_{0j})}\mathbb{1}\{j \neq i\}\Big)\mathbf{1}_p\mathbf{1}'_p\Big)\mathbf{P}_2^-\Big]
\end{aligned}
$$

Again, in the expression for $d_{q(\lambda_X)}^{(m)}$, $E_{(\theta_{-\lambda_X})}[\mathbf{f}(\mathbf{h}_i^{-1})]$ and $E_{(\theta_{-\lambda_X})}[\mathbf{f}(\mathbf{h}_i^{-1})\mathbf{f}(\mathbf{h}_i^{-1})']$ are unknown. Thus, we will make the following approximations

$$
E_{(\theta_{-\lambda_X})}[\mathbf{f}(\mathbf{h}_i^{-1})] \approx \boldsymbol{\mu}_{q(\mathbf{f})}(\mathbf{h}_i^{-1}) \text{ and } E_{(\theta_{-\lambda_X})}[\mathbf{f}(\mathbf{h}_i^{-1})\mathbf{f}(\mathbf{h}_i^{-1})'] \approx \boldsymbol{\Sigma}_{q(\mathbf{X}_i)}/N + \boldsymbol{\mu}_{q(\mathbf{f})}(\mathbf{h}_i^{-1})\boldsymbol{\mu}_{q(\mathbf{f})}(\mathbf{h}_i^{-1})'
$$

Due to these modifications, if noisy observations are observed the convergence properties of the adapted variational Bayes algorithm are not guaranteed to hold, and $log\ [f(\mathbf{Y}, \mathbf{w}; q)]$ cannot be monitored. However, you can proceed to monitor convergence for this model as follows. Taking advantage of the fact that functional smoothing converges more quickly than functional registration, fix the approximated unregistered functions, $\mathbf{X}_i$, $i = 1, \ldots, N$, after a small num-

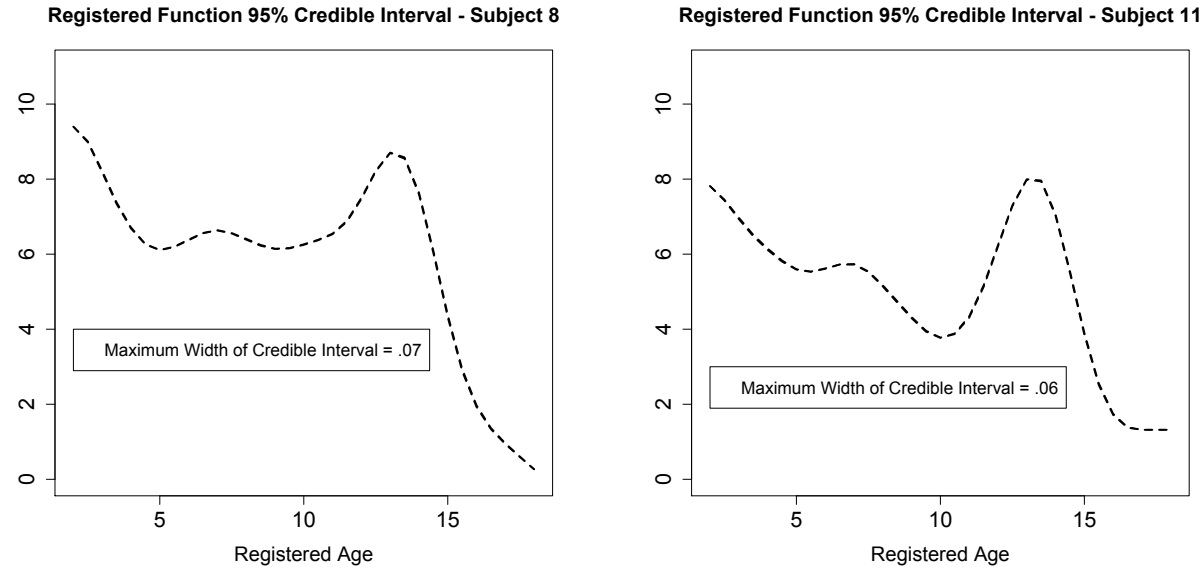

Figure 3.8: Examples of Credible Intervals for Noiseless Observations . These are two examples from the Boys Growth Velocity Data of the tight credible bands that result from registering functions that are pre-smoothed. In Figure 3.9, the top and lower right illustrations contain the credible intervals for these same observations when the noise process is included in the model.

ber of iterations and proceed as if they are known. Then, as in the model where the observations are recorded without noise, $log\ [f(\mathbf{X}, \mathbf{w}; q)]$ can be monitored.

3.5.3 The Berkeley Growth Data

We refer back to the Berkeley Boys Growth Velocity dataset from Section 3.3.1. In Section 3.3.1, these data were smoothed prior to registration. Here, we again consider these functions with the added assumption that they are corrupted by simulated mean zero iid Gaussian noise, where the true noise variance, σ_Y^2, is .25.

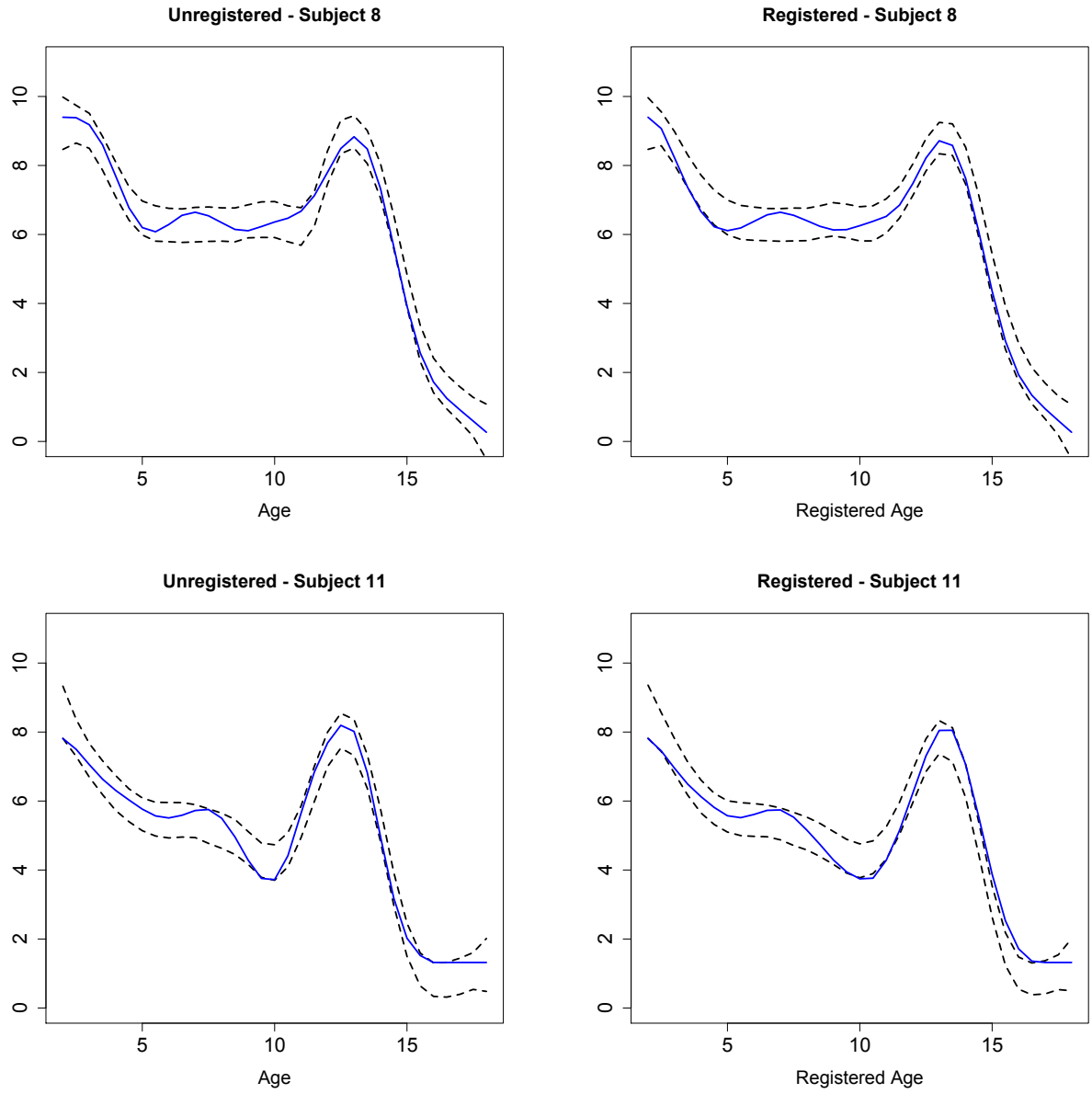

Figure 3.9: Examples of Credible Bands for the Unregistered and Registered Functions when the Noise Process is Included in the Model. **Top and Lower Left** 95% credible bands for the unregistered functions are plotted with the original noiseless functions for subjects 8 and 11. **Top and Lower Right** For subjects 8 and 11, 95% credible bands for the registered functions are plotted with the estimate of the registered 'true' functions.

While it is common in statistical analysis to perform preprocessing steps before applying a particular inference procedure, failing to account for the variability in parameter estimates due to the preprocessing step leads to overly narrow confidence (or credible) regions. In some cases, the effect may be fairly small, and not much is lost in this oversight. However, as we show here, there is the potential for the underestimation of variability to be substantial when uncertainty in the preprocessing steps is ignored.

In Section 3.3.2 is an illustration of how closely AVB and MCMC estimates of the registered functions adhere to one another. Not only do these estimates tend to be fairly similar when the functions are recorded without noise, but the uncertainty in these estimates is minimal. Figure 3.8 contains the credible bands for two of the 39 pre-smoothed Boys Growth Velocity Functions. These bands are so narrow the width between them cannot be seen. Keep in mind the posterior distributions of the registered functions are certainly multi-modal. These credible bands result from imposing the restriction that the mean value of the warping functions at each time point over the sample must equal that time point. Even with this restriction, the posterior distributions can be multi-modal. However, these narrow credible bands reflect that our estimates are in a highly probable area of the posterior distribution with minimal local variance. Figure 3.9 contains credible bands for both the unregistered and registered functions for the same two functions used in Figure 3.8 after noise has been added to the data and accounted for in the model. The variability due to noise is substantial. The solid line in all of the plots contains the noiseless version of these estimates (or observations in the case of the unregistered functions).

In addition to providing more accurate credible intervals, this model esti-

mates the noise variance to be .258 (actual noise variance is .25). This estimate is obtained using uninformative priors for both the noise variance, σ_Y^2 and the associated smoothing parameter λ_X.

This analysis illustrates how regularizing the data prior to statistical analysis for registration models severely limits inference for these models. If significant noise is present in the data it is prudent to account for the variability in the registration process due to the noise. Our proposed hierarchical model is one way to account for this variability.

CHAPTER 4

COMBINING FUNCTIONAL DATA REGISTRATION AND FACTOR ANALYSIS

4.1 Factor Analysis Models for Registration and Grouping

Here we extend our work on functional registration via Gaussian process models to allow for more flexible assumptions in the structure of the registered functions. Using the classical definition of functional registration, in Chapter 3 we propose a registration model designed to register functions have little variation from one functional direction. While appropriate for many statistical analyses, this registration model does not adequately register functions in which there are more than one primary direction of variation in the registered functions. As we will show in Section 4.2, other registration methods based on this traditional definition of registration also tend to perform poorly when the registered functions are composed of more than one primary direction of variation.

In Chapter 3, we established that under the assumption that the registered functions do not vary significantly from one primary functional direction, the following data distribution is appropriate to register functions $X_i(t)$, $i = 1, \ldots, N$.

$$X_i(h_i(t)) \mid z_{0i}, z_{1i}, f_1(t) \sim GP(z_{0i} + z_{1i} f_1(t), \gamma_1^{-1}\Sigma(s,t)) \quad s,t \in \mathcal{T} \tag{4.1}$$

where $X_i(h_i(t))$ is $X_i(t)$ registered under the warping $h_i(t)$. The above covariance function, $\gamma_1^{-1}\Sigma(s,t)$, penalizes all variance from a scaling and vertical shifting of the primary functional direction, $f_1(t)$. In these models we define γ_1 as a registration parameter that determines the severity of this penalty. This registration parameter is balanced by a penalty on the warping functions, $h_i(t), i = 1, \ldots, N$

that penalizes distance from the identity warping.

It is natural to extend this initial model to

$$X_i(h_i(t)) \mid z_{0i}, z_{1i}, f_1(t), z_{2i}, f_2(t) \sim GP(z_{0i} + z_{1i}f_1(t) + z_{2i}f_2(t), \gamma_1^{-1}\Sigma(s,t)) \quad s,t \in \mathcal{T} \tag{4.2}$$

However, this distribution penalizes variation from the first and second functional directions (factors), $f_1(t)$ and $f_2(t)$, equally. For most data scenarios, variation in one of the factors will exceed variation in the other factor. Accounting for this discrepancy in the statistical model for the registered functions not only provides a better registration, but also creates an identifiable relationship between the two factors. We will thus proceed with the following distribution on the registered functions.

$$X_i(h_i(t)) \mid z_{0i}, z_{1i}, f_1(t), z_{2i}, f_2(t) \sim GP(z_{0i} + z_{1i}f_1(t) + \frac{\gamma_2}{\gamma_1+\gamma_2}z_{2i}f_2(t), (\gamma_1+\gamma_2)^{-1}\Sigma(s,t)) \quad s,t \in \mathcal{T}$$

Before establishing the basis for the distribution specified above , we note here, as is common with functional data, we assume observations of each unregistered function, $X_i(t)$ are observed over a finite number of equally spaced time points, $\mathbf{t} = (t_1, \ldots, t_p)'$. Thus, given the above model, in practice we will proceed by using finite approximations to each functional distribution. In Section 2.1.2, we establish some theoretical properties of these types of approximations. The following finite-dimensional distribution is used in the final model in lieu of its infinite dimensional counterpart above. For $\mathbf{X}_i(\mathbf{h}_i) = (X_i(h_i(t_1)), \ldots, X_i(h_i(t_p)))'$, $i = 1, \ldots, N$

$$\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}_1, z_{2i}, \mathbf{f}_2 \quad \sim \quad N_p(z_{0i}\mathbf{1} + z_{1i}\mathbf{f} + \frac{\gamma_2}{\gamma_1+\gamma_2}z_{2i}\mathbf{f}_2, (\gamma_1+\gamma_2)^{-1}\mathbf{\Sigma}) \tag{4.3}$$

For Gaussian data, it is easy to see the justification for constructing precision matrices that ensure desirable properties in the model. For example, if we as-

sume the estimates of approximated functions, $\mathbf{X}_i(\mathbf{h}_i)$, are smooth, the precision matrix for the distribution of $\mathbf{X}_i(\mathbf{h}_i)$ can be at least partially constructed by using the square of a finite second difference matrix. Assuming the second finite difference penalty matrix is $\mathbf{L}_2$ and the distribution of $\mathbf{X}_i(\mathbf{h}_i)$ is

$$\mathbf{X}_i(\mathbf{h}_i) \mid \eta, \lambda \quad \sim \quad N_p(\mathbf{0}, (\eta\mathbf{P} + \lambda\mathbf{L}_2'\mathbf{L}_2)^{-1})$$

where $\mathbf{P}$ is defined to establish a proper covariance matrix as $\mathbf{L}_2'\mathbf{L}_2$ is singular, then,

$$\mathbf{X}_i(\mathbf{h}_i) \mid \eta, \lambda \quad \propto \quad exp[-\frac{1}{2}(\mathbf{X}_i(\mathbf{h}_i)'\lambda\mathbf{L}_2'\mathbf{L}_2\mathbf{X}_i(\mathbf{h}_i))]$$

If we considered this as a likelihood function, the expression within the exponential is largest when the square of the summed second derivatives (point wise) of $\mathbf{X}_i(\mathbf{h}_i)$ are small indicating the smoother $\mathbf{X}_i(\mathbf{h}_i)$ is, the more probable it becomes. From a Bayesian perspective, assuming this type of prior (or data) distribution, assures smoothness in the posterior distribution of $\mathbf{X}_i(\mathbf{h}_i)$, where the level of smoothness is defined by λ.

Now switching back to our model in (4.3). The precision matrix, $\mathbf{\Sigma}^{-1}$, as defined in Section 3.1, is designed to penalize all variation from a given mean function. In our model, we would like to penalize variation from the first and second factors using two separate registration parameters, γ_1 and γ_2, so that the prior for $\mathbf{X}_i(\mathbf{h}_i)$ has the following property

$$\begin{aligned}
&\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}_1, z_{2i}, \mathbf{f}_2 \propto \\
&\quad exp \quad [-\frac{1}{2}((\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1))'\gamma_1\mathbf{\Sigma}^{-1}(\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1)))] * \\
&\quad exp \quad [-\frac{1}{2}((\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1 + z_{2i}\mathbf{f}_2))'\gamma_2\mathbf{\Sigma}^{-1}(\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1 + z_{2i}\mathbf{f}_2)))]
\end{aligned}$$

After rearranging terms, and ignoring everything constant in $\mathbf{X}_i(\mathbf{h}_i)$, this criterion results in prior distribution (4.3) for the registered functions.

The full data and prior distributions for the factor analysis registration model assuming unregistered functions $X_i(t)$, have been observed over $\mathbf{t} = (t_1, \ldots, t_p)'$ are

$$
\begin{aligned}
\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}_1, z_{2i}, \mathbf{f}_2 &\sim N_p(z_{0i}\mathbf{1} + z_{1i}\mathbf{f}_1 + \frac{\gamma_2}{\gamma_1+\gamma_2} z_{2i}\mathbf{f}_2, (\gamma_1+\gamma_2)^{-1}\mathbf{\Sigma}) \quad i = 1, \ldots, N \\
\mathbf{h}_i(t_j) &= t_1 + \sum_{k=2}^{j} (t_k - t_{k-1}) e^{w_i(t_{k-1})} \quad i = 1, \ldots, N \quad j = 1, \ldots, p \\
\mathbf{w}_i \mid \lambda_w &\propto N_{p-1}(\mathbf{0}, \gamma_w^{-1}\mathbf{\Sigma} + \lambda_w^{-1}\mathbf{P}_w) \mathbb{1}\{t_1 + \sum_{k=2}^{p} (t_k - t_{k-1}) e^{w_i(t_{k-1})} = t_p\} \quad i = 1, \ldots, N \\
z_{0i} \mid \sigma^2_{z0} &\sim N(0, \sigma^2_{z0}) \quad i = 1, \ldots, (N-1) \quad z_{0N} = -\sum_{i=1}^{N-1} z_{0i} \\
\sigma^2_{z0} &\sim IG(a, b) \\
z_{1i} \mid \sigma^2_{z1} &\sim N(1, \sigma^2_{z1}) \quad i = 1, \ldots, N \\
\sigma^2_{z1} &\sim IG(a, b) \\
z_{2i} \mid \sigma^2_{z2} &\sim N(0, \sigma^2_{z2}) \quad i = 1, \ldots, N \\
\sigma^2_{z2} &\sim IG(a, b) \\
\mathbf{f}_1 \mid \eta_f, \lambda_f &\sim N_p(0, \mathbf{\Sigma}_f) \\
\mathbf{f}_2 \mid \eta_f, \lambda_f &\sim N_p(0, \mathbf{\Sigma}_f) \\
\mathbf{\Sigma}_f &= \eta_f^{-1}\mathbf{P}_1 + \lambda_f^{-1}\mathbf{P}_2 \\
\eta_f &\sim G(c, d) \\
\lambda_f &\sim G(c, d)
\end{aligned}
$$

In this model, a, b, c, and d are hyper-parameters defining uninformative priors on the variance components and smoothing parameters. The matrix, $\mathbf{P}_2$ is designed to penalize smoothing in the factors, $\mathbf{f}_1$ and $\mathbf{f}_2$, with corresponding

smoothing parameter, λ_f. $\mathbf{P}_1$ is defined to establish a positive definite covariance matrix for the distributions of the two factors. The corresponding parameter, η_f, must just be large enough to stabilize this matrix. Note, η_f and λ_f are considered as additional unknown parameters to be estimated through the model.

In addition to allowing more flexibility in the shape of the registered functions, a bi-product of this analysis is the estimation of the two functional directions, $f_1(t)$ and $f_2(t)$, and the associated weights of these two factors for each function, z_{1i} and z_{2i}, $i = 1, \ldots, N$, respectively. These factors tend to have a more interpretable shape than principal components, and estimating the weights for each function provides a way to group registered functions.

As is typical with hierarchical models, all parameters can be estimated using MCMC samples from the joint posterior distribution. However, obtaining these samples in high-dimensional models can be expensive and time-consuming. In Section 3.2.1 we define and establish convergence properties for an adapted version of variational Bayes that can also be utilized here. Appendix C contain all of the model specifications, full-conditionals for a MCMC sampler, and details of the adapted variational Bayes algorithm.

4.2 Comparison to Current Methods

One of the best registration methods currently available is that proposed by Srivistava, et.al. [43]. The authors build a registration model based on the Fisher-Rao Riemannian metric that is superior to many previously considered algorithms (F-R method). Further details on the F-R registration method can be found in Section 3.3.1.

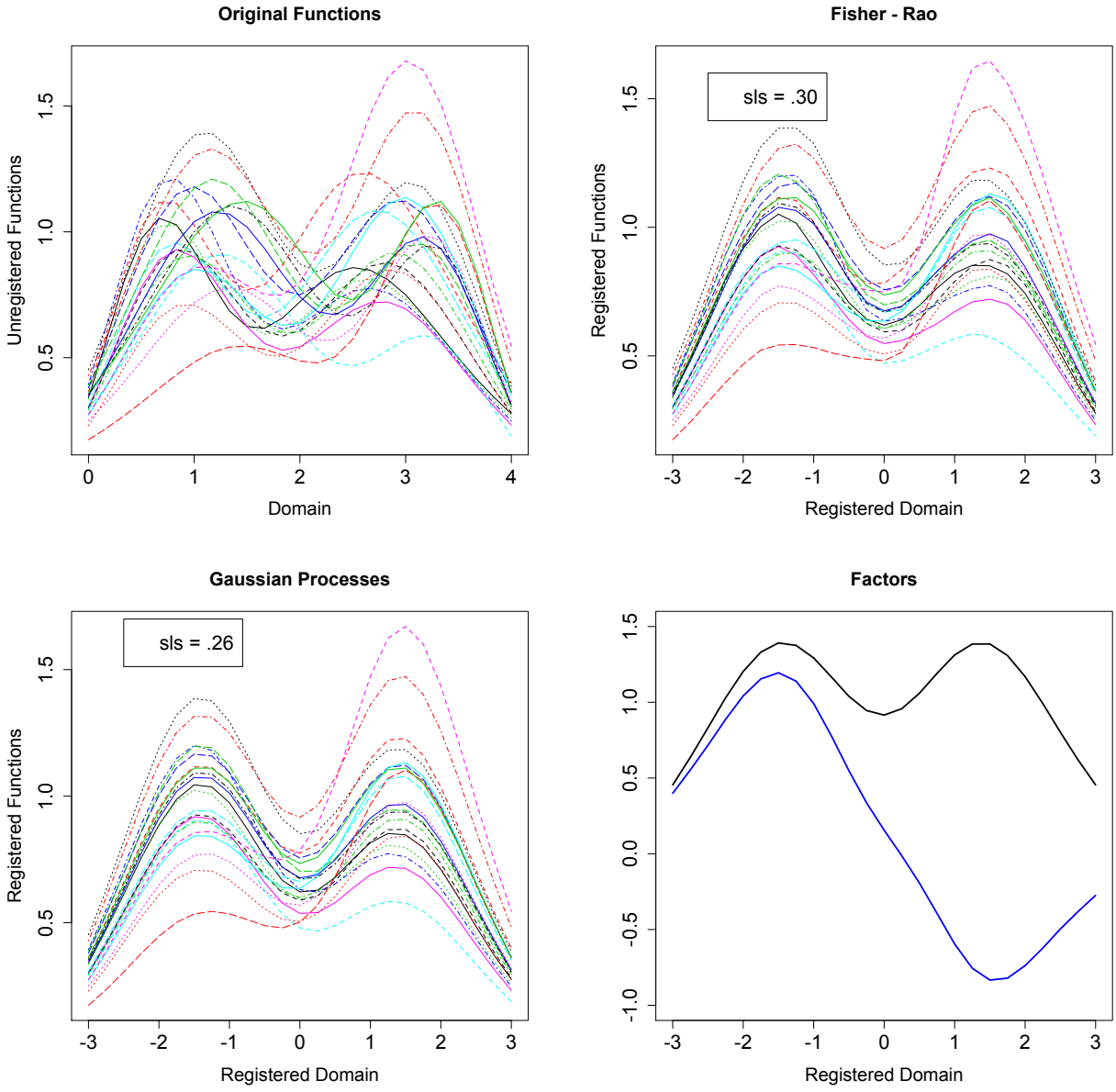

Figure 4.1: First Simulated Data Set. **Top Left** Original unregistered functions. **Top Right** Functions registered by F-R (R package 'fdasrvf'). **Lower Left** Functions registered by the FA model. **Lower Right** Estimated factors $\mathbf{f}_1$ and $\mathbf{f}_2$.

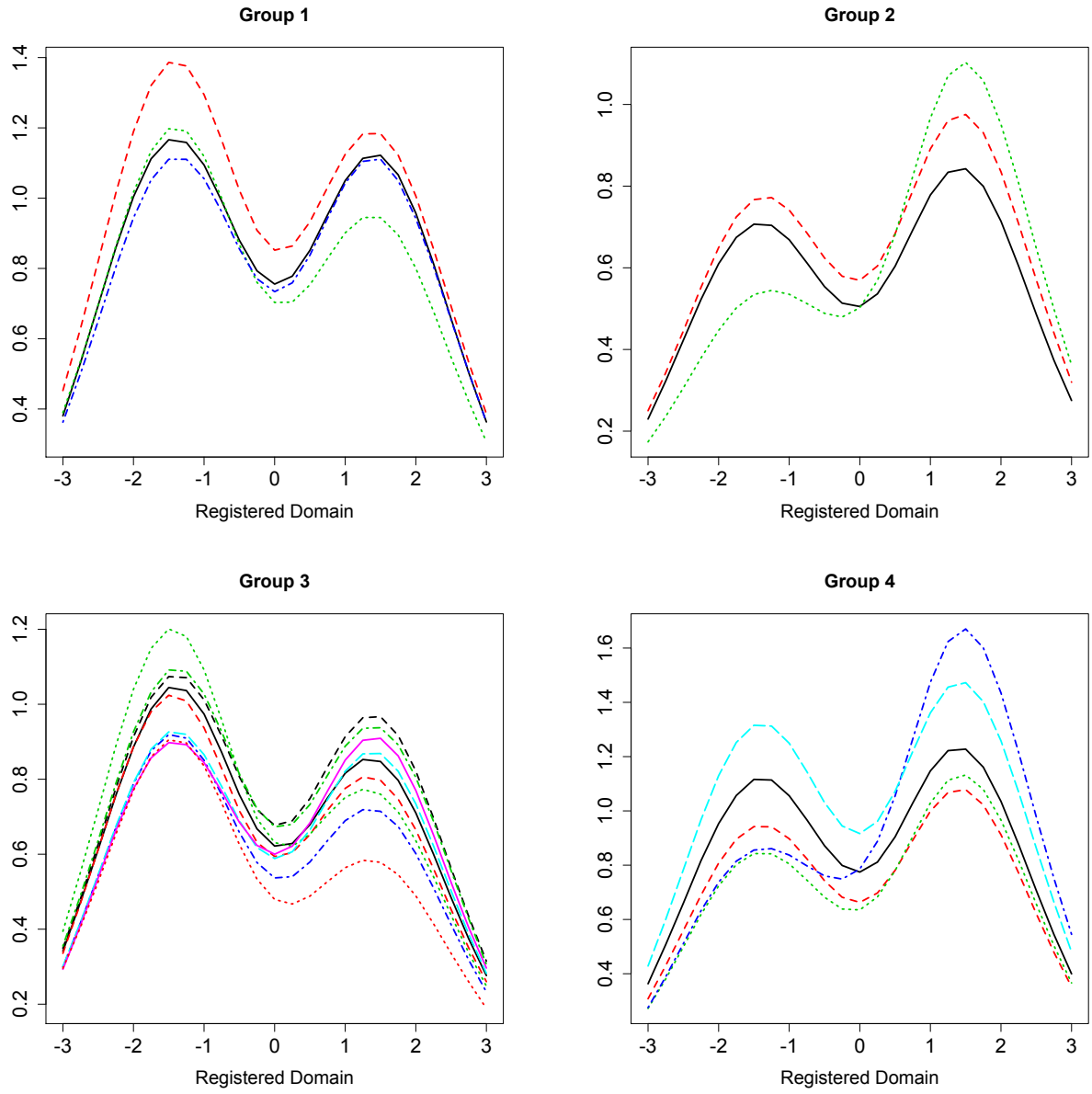

Figure 4.2: Four groups determined by the centered weights, $\tilde{\mathbf{z}}_1$ and $\tilde{\mathbf{z}}_2$. **Top Left** $\{X_i(h_i(t)) : \tilde{z}_{1i} > 0, \tilde{z}_{2i} > 0\}$. **Top Right** $\{X_i(h_i(t)) : \tilde{z}_{1i} < 0, \tilde{z}_{2i} < 0\}$ **Lower Left** $\{X_i(h_i(t)) : \tilde{z}_{1i} < 0, \tilde{z}_{2i} > 0\}$ **Lower Right** $\{X_i(h_i(t)) : \tilde{z}_{1i} > 0, \tilde{z}_{2i} < 0\}$

In Chapter 3, we present registration results similar to the F-R method using a Gaussian process model (GP) (3.4). The extension of this model proposed here improves on the F-R method for certain types of data. Here, we will compare the registration results of F-R and of our GP model using two simulated data sets. Registered functions under both models are compared using the Sobolev Least Squares criterion (3.9) where lower values correspond to better alignment.

First Simulated Data Set The 21 unregistered functions are simulated using the algorithm originally proposed by Kneip and Ramsay [22] where the authors also consider registration in the context of multiple directions of functional variation. The registered functions $X_i(h_i(t))$, $i = 1, \ldots, 21$, are defined as $X_i(h_i(t)) = c_{1i}e^{-.5(t-1.5)^2} + c_{2i}e^{-.5(t+1.5)^2}$, $t \in [-3, 3]$ where c_{1i} and c_{2i} are iid $N(1, .25^2)$. These functions are then warped so that $h_i(t) = 6(\frac{e^{a_i(t+3)/6}-1}{e^{a_i}-1}) - 3$ if $a_i \neq 0$, where $a_i, i = 1, \ldots, 21$ are equally spaced between -1 and 1. If $a_i = 0$, $h_i(t) = t$.

Data simulated in the same way are also registered using the F-R method in Srivistava, et.al. [43]. Here we again use their method to register the simulated unregistered functions for comparison purposes. In Figure 4.1 plots of the simulated unregistered functions and the functions registered using both the F-R algorithm and the proposed GP model. Both methods achieve a high degree of alignment with the GP model performing slightly better using the *sls* criterion. The lower left frame of Figure 4.1 contains the two estimated factors to which these data are registered.

While the GP model performs similarly to F-R in this example, the added benefit of using the GP model is that the registered functions can be grouped according to their associated weights, $\mathbf{z}_1$ and $\mathbf{z}_2$ on each of the factors, $\mathbf{f}_1$ and $\mathbf{f}_2$. Figure 4.2 are four groups of estimated registered functions, based on the

set of estimated centered weights $\tilde{\mathbf{z}}_1$ and $\tilde{\mathbf{z}}_2$, where all functions whose centered weights lie in the same quadrant are grouped together.

Second Simulated Data Set Here we consider data with features that are not aligned well using traditional definitions of registration. Each of the 20 simulated registered functions is composed of a linear combination of two factors which is then subjected to a random warping to obtain a simulated unregistered function. The factors, $\mathbf{f}_1$ and $\mathbf{f}_2$ from which these data are simulated are found in Figure 4.3.

The alignment of these functions using the GP model is again compared to that obtained by F-R. For this example, the quality of alignment is best assessed by using the Sobolev Least Squares criterion separately for each of two groups of functions. Group 1 consists of functions for which $\hat{z}_{2i} > 0$. The second group is characterized by functions for which $\hat{z}_{2i} < 0$. The final *sls* value is the sum of the *sls* values for the two groups.

In Figure 4.3 are plots of the simulated unregistered functions, the functions registered by F-R, and the functions registered by GP. Not only is the *sls* value lower for the GP model, visually it is apparent that functions registered by the GP model are better aligned. In this example the estimated factors closely resemble the original factors from which the data are simulated. These can be seen in Figure 4.4. Also, in Figure 4.4 are three groups of registered functions determined only by classifying the estimated weights on the second factor.

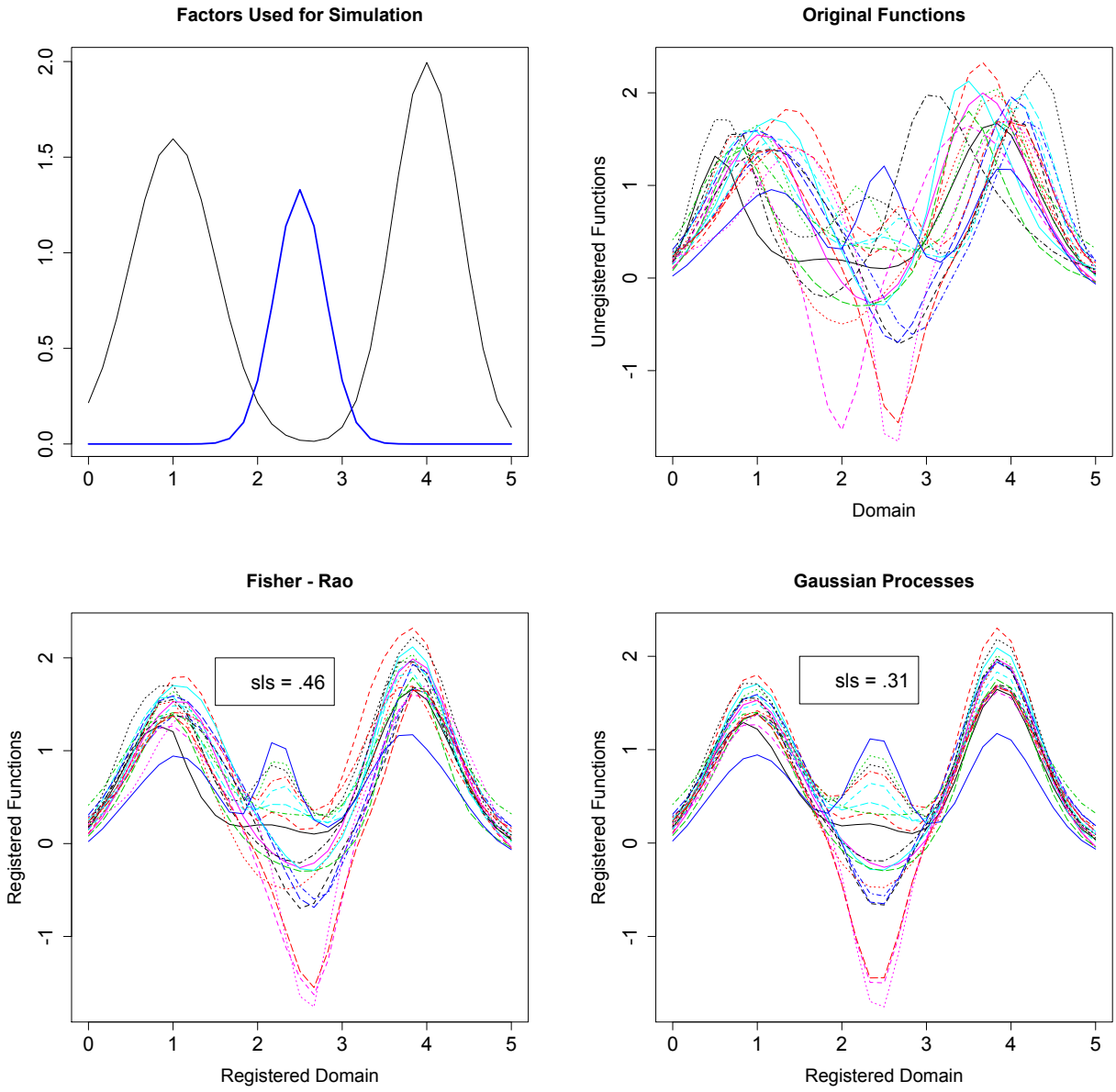

Figure 4.3: Second Simulated Data Set. **Top Left** The two factors used to simulate data before warping. **Top Right** Simulated unregistered functions. **Lower Left** Functions registered by F-R (R package 'fdasrvf'). **Lower Right** Functions registered by the GP model.

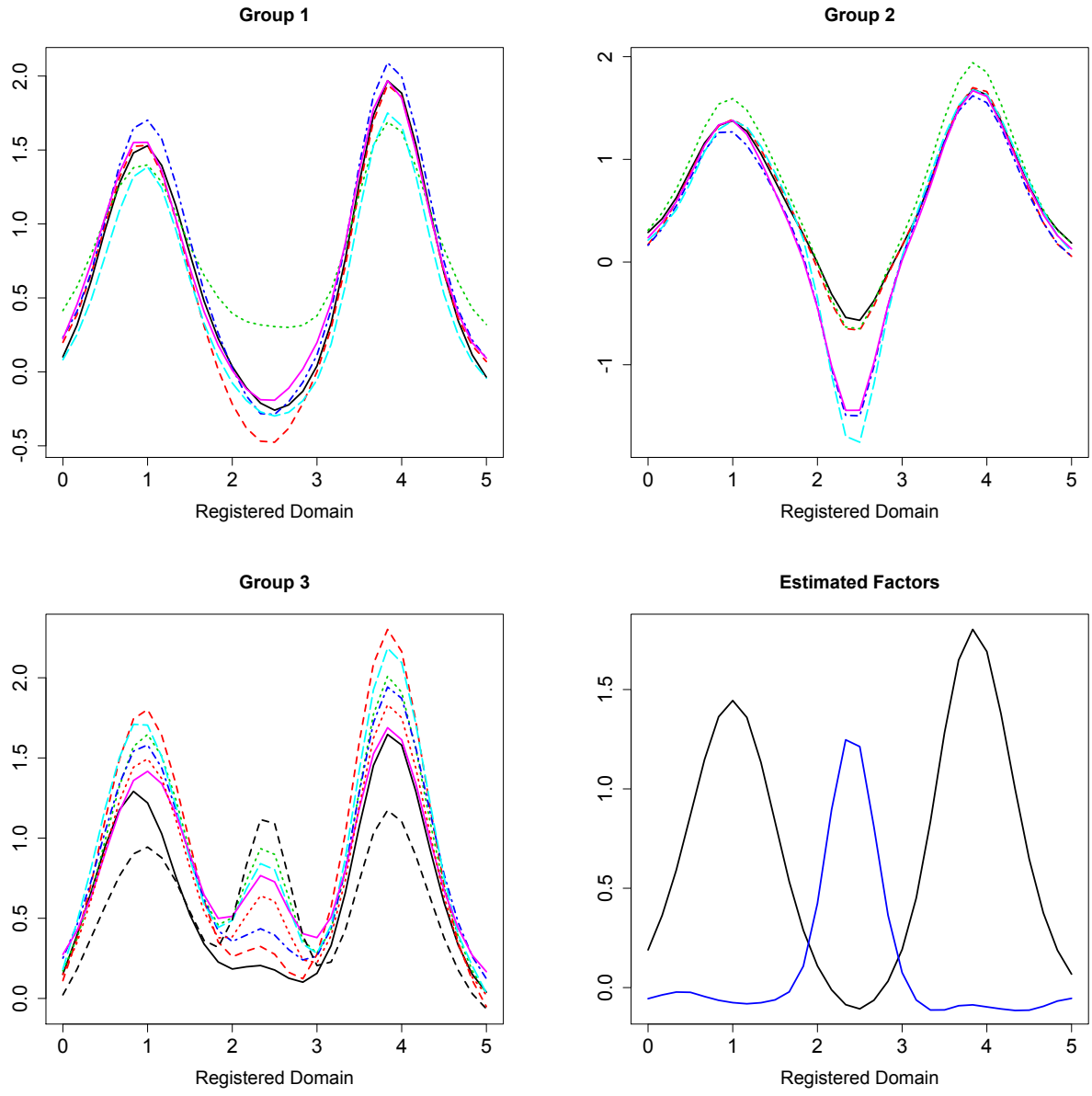

Figure 4.4: Three groups determined by the estimated weights on the second factor, $\mathbf{z}_2$. **Top Left** $\{X_i(h_i(t)) : \hat{z}_{2i} \in [-.1, .1]\}$. **Top Right** $\{X_i(h_i(t)) : \hat{z}_{2i} < -.1\}$ **Lower Left** $\{X_i(h_i(t)) : \hat{z}_{2i} > .1\}$ **Lower Right** Estimated factors, $\hat{\mathbf{f}}_1$ and $\hat{\mathbf{f}}_2$, determined by the GP model.

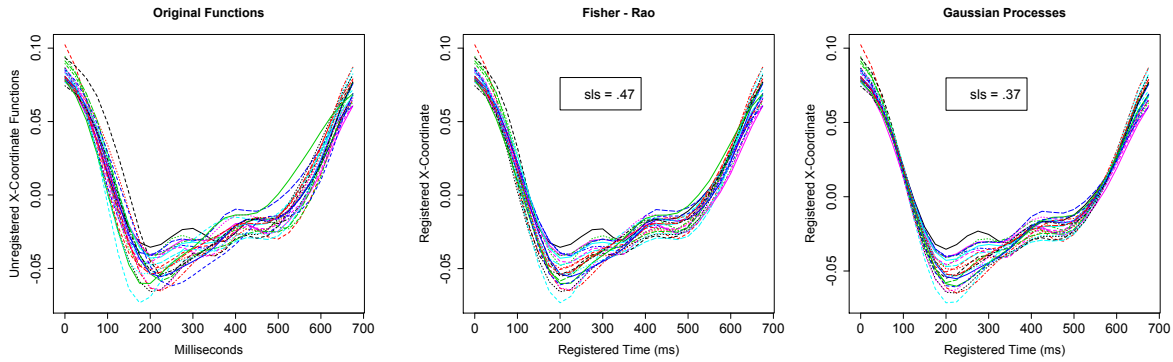

Figure 4.5: Juggling Data. **Top Left** Original unregistered functions. **Top Right** Functions registered by F-R (R package 'fdasrvf'). **Lower Left** Functions registered by the FA model. **Lower Right** Estimated factors, $\hat{\mathbf{f}}_1$ and $\hat{\mathbf{f}}_2$, determined by the GP model.

4.3 The Juggling Data: Registration and Grouping

The juggling data consist of three different functional data sets obtained by recording the finger position of Dr. Michael Newton (Biostatistics, University of Wisconsin-Madison) as he juggles. These data were collected in collaboration with Dr. James Ramsay (Psychology, McGill University), Dr. David Ostry (Psychology, McGill University), and Dr. Paul Gribble (Psychology, University of Western Ontario). As Dr. Newton juggled the following were recorded: 1) the horizontal position of the right forefinger in the frontal plane, 2)the horizontal position of the right forefinger in the sagittal plane, and 3) the vertical position of the right forefinger. For this data analysis, the first functional data set of the horizontal position of the right forefinger in the frontal plane is used to demonstrate functional data registration and grouping using our Gaussian process model. Additional information on this data set can be found in Ramsay and Silverman [32].

Description of the Juggling Data The first data set consists of ten functional

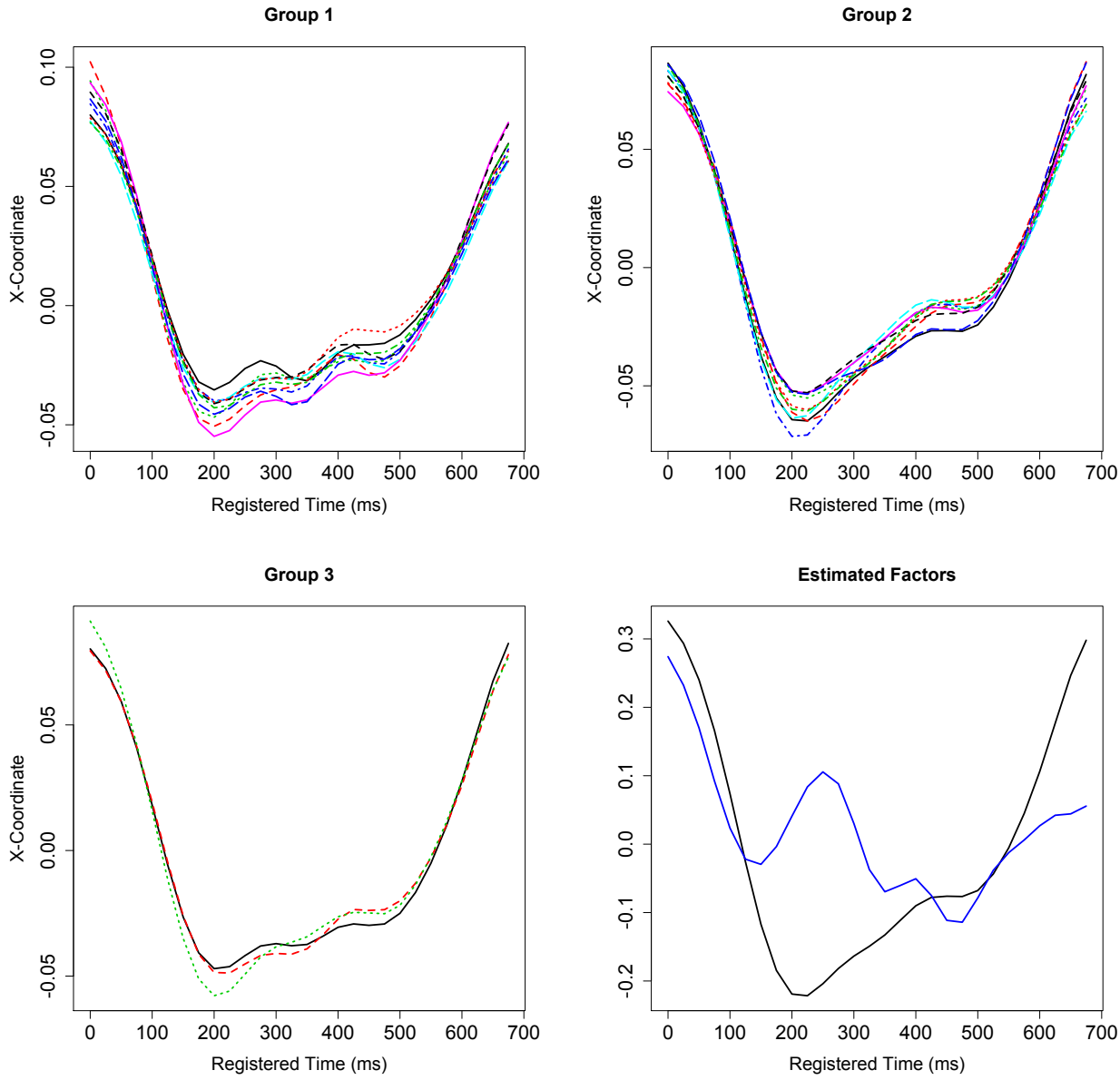

Figure 4.6: Three groups determined by the estimated centered and scaled weights on the second factor, $\tilde{\mathbf{z}}_2$. **Top Left** $\{X_i(h_i(t)) : \tilde{z}_{2i} > .1\}$. **Top Right** $\{X_i(h_i(t)) : \tilde{z}_{2i} < -.1\}$ **Lower Left** $\{X_i(h_i(t)) : \tilde{z}_{2i} \in [-.1, .1]\}$

observations each ranging from 11 to 13 juggling cycles. For this analysis, our observations consist of individual cycles. Each functional observation begins at the apex of each cycle that corresponds to the position of the juggler's right forefinger immediately preceding the start of a cycle. From here, each function takes a sharp dip as the juggler brings down the ball and subsequently snaps the ball upward as it is released. Of approximately 120 cycles available, we selected 23 that met the following criteria: 1) each cycle was of the same length, and 2) each of the functions followed one of two distinct trajectories. Each observation was 675 milliseconds in length where the original data are recorded in 5 millisecond intervals. Thinning the data does not significantly alter its shape, and the final data contains 27 records per functional observation (cycle) taken every 25 milliseconds. Additionally, extra warping was introduced in each of the 23 functions to clearly illustrate the registration capabilities of this model.

The goal of this analysis is two-fold. The first aim is to align the prominent features in these 23 cycles in conjunction with estimating the two primary factors of which these data are composed. Secondly, using the estimated weights, $\hat{z}_{1i}$ and $\hat{z}_{2i}$, classify these functions into distinct groups. Figure 4.5 contains plots of the unregistered functions, the functions registered by F-R, and the functions registered by GP. Here again, based on the *sls* criterion, the GP model provides a better function alignment than F-R. The estimated registered functions are split into three groups in Figure 4.6. Similar to the second simulated data set found in Section 4.2, these groups are based solely on the estimated weights on the second factor. In Figure 4.6 it can be seen that the first estimated factor strongly resembles the X-coordinate over time of the juggling cycles found in Group 2. The cycles in this group are associated with a small weight on the second factor. Group 1 contains juggling cycles for which variation in the X-coordinate of each

cycle has a strong weighting on the second factor. The second estimated factor gives a clear illustration of where extra movement in the right forefinger can be found in the cycles in Group 1 as compared to those in Group 2. Group 3 contains functions for which the effect of the second factor is more subtle.

CHAPTER 5

SUMMARY OF FINDINGS, DISCUSSION, AND FUTURE WORK

5.1 Summary of Findings

The major contributions of this dissertation lie in the use of random effect models in a Bayesian environment for smooth functional data characterized by a Gaussian process. Within this framework, the following areas of inference in functional data analysis are addressed: non-parametric covariance estimation, functional data smoothing, smoothing parameter selection, functional linear regression, functional data registration, and simultaneous factor analysis and registration for functional data. The wide scope of this work emphasizes the flexibility afforded in using a hierarchical Bayesian model for functional data analysis. Furthermore, this work addresses the computational challenges associated with high-dimensional models through the development of an adapted variational Bayes algorithm.

All functional data in these models are characterized by either a Gaussian process or a functional inverse-Wishart distribution. These distributions provide a non-parametric alternative to the use of basis systems to model functional data. However, these distributions are not tractable to work with directly. In this dissertation, the theoretical properties of using finite dimensional distributions to approximate these infinite dimensional distributions are provided, and it is shown that functional estimates at time points where data are not observed obtained by simple linear (or bi-linear) interpolation have nice properties.

The theoretical justification for using a finite approximation to a functional

inverse-Wishart distribution is crucial for the novel work in this dissertation in the area of covariance function estimation. Using a finite approximation to the functional inverse-Wishart distribution allows the covariance function to be modeled as a random effect. In this Bayesian environment inference for the covariance function in addition to all other random effects can be performed through the posterior distribution of these parameters. There are two significant benefits to modeling the covariance function as a random effect: 1) the variability in the covariance function estimate is easily quantified through the posterior distribution of this parameter, and 2) in contrast to functional data analysis where the covariance function is estimated prior to analysis and considered as a known hyper-parameter, the variability associated with estimating the covariance function is captured in this model and results in credible intervals with better coverage properties for all unknown parameters.

Many of the models in this dissertation take into account that functional data are often recorded noisily. Priors that include smoothing information are shown to be effective in estimating the underlying noiseless functional data. In this work, it is shown that smoothing parameters can be effectively automatically selected through the hierarchical model. These parameters are considered as additional unknown parameters with uninformative prior distributions. This approach to smoothing parameter selection avoids traditional pitfalls associated with cross-validation procedures or simply choosing these parameters *ad-hoc*.

The next extension of random effect models for functional data analysis presented in this dissertation is in the area of functional data registration. Here, again, it is shown how the use of an informative precision matrix in the hierarchical model provides desirable properties in function estimates. This model

for registering functional data is shown to be as good or better than current registration methods. This model distinguishes itself from other approaches to functional data registration in the use of an identifiable warping function. A functional prediction model for the warping function, registered function, and unregistered function are presented for the first time in this dissertation.

The registration model is further extended to a combined factor analysis and registration model. In this dissertation, it is shown that for data that vary in more than one registered functional direction, the factor analysis and registration model presented here provides significant improvements in registration over other current methods. Furthermore the two primary directions of variation in the registered functions are estimated in the model in conjunction with the estimated weights for each observation. This additional information extends the inferential capabilities of a traditional registration algorithm.

Finally, this dissertation addresses the computational issues associated with inference in high-dimensional Bayesian models by developing and establishing properties for the adapted variational Bayes algorithm as an approximation to or as an initialization method for MCMC sampling. This algorithm offers significant savings in computational costs in both the registration and combined factor analysis and registration models. Here, it is shown that estimates from the adapted variational Bayes algorithm tend to strongly coincide with their MCMC sampling counterparts. This algorithm is particularly useful for determining registration parameters and warping penalties that can then be used in the final MCMC sampling scheme if desired.

5.2 Discussion and Future Work

The primary advantage of the hierarchical Bayes models discussed in this dissertation is two-fold. First of all, these models provide a straight-forward set-up for otherwise complicated statistical procedures. The hierarchy in these models provides transparency in the mechanisms that drive smoothing and registration through the use of precision matrices that penalize undesirable properties in function estimates. Secondly, these models offer a unified approach to inference where all unknown parameters are estimated in one model (except for the registration parameters and warping penalties which must be chosen to provide a desired level of alignment) . Avoiding pre-processing steps is crucial for adequately quantifying variance in the posterior distributions for all unknown parameters in hierarchical Bayesian models.

The main drawback to these types of models is the computational costs. While these costs can be significantly reduced using the adapted variational Bayes algorithm for both registration models, this algorithm cannot be applied when the covariance function is considered as a random effect. The time-consuming nature of using a MCMC sampler in these high-dimensional models is generally best done with the use of high performance computers. This may restrict the accessibility of these models for general use. However, this obstacle lessons over time as significant advances in computational speed continue to develop. Future work lies in the area of using alternative sampling schemes to a Gibbs sampler that are more efficient. In particular Calderhead et. al. [6] suggests that population MCMC can be employed to allow both global and local movement throughout the parameter space for a more efficient sampler.

Another area of future work concerns the use of uninformative inverse-Gamma or Gamma priors for variance components in these models. These priors have the property of being conditionally conjugate with the data distribution and have been used for convenience. However, work by Gelman [10] suggests that these priors tend to be informative when the variance components are close to zero or when little data is available. In general, these priors seemed adequate in these models. However, there is some evidence that the estimate of the variance parameter for the noise in the registration model for the Berkeley Boys Growth Velocity data analysis in Section 3.5.3 is slightly inflated. Ideally, these priors should be replaced by weakly informative uniform priors on the square root of the variance component as suggested by Gelman [10] in his paper.

Finally, future work also should focus on methodical approaches to selecting registration parameters and warping penalties in these models. These are currently chosen by simply looking at the registered function estimates after the adapted variational Bayes algorithm has been run for a couple of iterations. Since, warping functions are estimated through maximizing the current iteration's likelihood function in this algorithm, estimates for these functions converge in a small number of iterations. One or two iterations is often enough to determine whether the registration and warping parameters are set at levels that will produce desirable registration results. However, even one or two iterations of this algorithm can take a significant amount of time to run. One option may be to initially analyze only a subset of the original data until these parameters are established. Another aspect that could be addressed is that the ratio of the registration parameter to the warping penalty is likely to be more important in these models than the actual values of these parameters. Similarly, in the factor analysis model, the ratio of the two registration penalties is likely

to be more important than there actual values. Determining efficient heuristics for selecting these parameters could add significant value to these models.

APPENDIX A

APPENDIX TO CHAPTER 2

A.1 Smoothing Parameter Selection

The expected draw of $\lambda_1^{(m+1)}$ is determined using the full-conditional distribution for λ_1 given in Appendix A.2. In (A.1) below, $\boldsymbol{\Sigma}_X^{-1(m+1)}$ has been replaced by its expectation in order to examine the relationship between a new draw of the smoothing parameter and the previous draw of this parameter. The full-conditional distribution from which this expectation has been derived can be found in Appendix A.2. We will assume for this exhibition that the hyperparameters, a and b, associated with the inverse Gamma prior defined for λ_1 are sufficiently small that they can effectively be ignored.

$$
\begin{aligned}
E(\lambda_1^{(m+1)}) &= \frac{tr(\mathbf{P}_2\boldsymbol{\Sigma}_X^{-1(m+1)})}{(p+1)(p-2)-2} \\
&\approx \frac{tr(\mathbf{P}_2 E(\boldsymbol{\Sigma}_X^{-1(m+1)}))}{(p+1)(p-2)-2} \qquad \text{(A.1)} \\
&= \frac{(p+1+N)tr(\mathbf{P}_2(\eta_1^{-1(m)}\mathbf{P}_1 + \lambda_1^{-1(m)}\mathbf{P}_2 + \sum_{i=1}^{N}(\mathbf{X}_i-\boldsymbol{\mu})(\mathbf{X}_i-\boldsymbol{\mu})')^{-1})}{(p+1)(p-2)-2} \\
&= \frac{(p+1+N)tr((\sum_{j=3}^{p}\kappa_j^{-1}\boldsymbol{\nu}_j\boldsymbol{\nu}_j')(\sum_{j=1}^{2}\frac{\eta_1^{(m)}}{1+\eta_1^{(m)}\sum_{i=1}^{N}c_{ij}^{2(m)}}\boldsymbol{\nu}_j\boldsymbol{\nu}_j' + \sum_{j=3}^{p}\frac{\lambda_1^{(m)}\kappa_j}{1+\lambda_1^{(m)}\kappa_j\sum_{i=1}^{N}c_{ij}^{2(m)}}\boldsymbol{\nu}_j\boldsymbol{\nu}_j'))}{(p+1)(p-2)-2} \\
&= \frac{(p+1+N)tr(\sum_{j=3}^{p}\frac{\lambda_1^{(m)}}{1+\lambda_1^{(m)}\kappa_j\sum_{i=1}^{N}c_{ij}^{2(m)}}\boldsymbol{\nu}_j\boldsymbol{\nu}_j')}{(p+1)(p-2)-2} \\
&\approx \lambda_1^{(m)}\Big(\frac{p+1+N}{p+1}\Big)\frac{1}{p-2}\sum_{j=3}^{p}\frac{1}{1+\lambda_1^{(m)}\kappa_j\sum_{i=1}^{N}c_{ij}^{2(m)}} \qquad \text{(A.2)}
\end{aligned}
$$

Here, $\{\boldsymbol{\nu}_j : j = 1 \dots p\}$ are the eigenvectors of the inverse scale matrix, $\eta_1\mathbf{P}_1^- + \lambda_1\mathbf{P}_2^-$, where $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ penalize linear and constant variation while

$\{\nu_j : j = 3 \dots p\}$ penalize curvature. $\{\eta_1, \eta_1, \{\lambda_1 \kappa_j : j = 3 \dots p\}\}$ are the corresponding eigenvalues of the inverse scale matrix. Furthermore, for each j, $\sum_{i=1}^{N} c_{ij}^{2(m)}$ represents the variation present in the latent functions in the jth direction in iteration m that is determined by representing each centered approximation to a latent function as a linear combination of the eigenvectors of the penalty matrix. The approximation in (A.2) is due to the omission of a factor of $\left(1 - \frac{2}{(p+1)(p-2)}\right)^{-1}$.

A.2 Distributional Assumptions

Below, in detail, are the joint data distributions, prior distributions, and full conditional distributions for the models discussed in Chapter 2. The first section describes the basic model for smoothing, estimating, and characterizing latent functional data. The next section expands this model to encompass functional linear regression. The third section looks at how to adjust the Gibbs sampler to account for missing observations.

A.2.1 Estimating Latent Functional Data

As discussed in Section 2.1, the initial assumption of this model is that we are interested in the functional data, $X_i(t), i = 1, \dots, N$, modeled by a Gaussian process, for which we only have noisy observations of each function at a given set of time points, $t_j, j = 1, \dots, p$. Observations, $Y_i(t_j), i = 1, \dots, N, j = 1, \dots, p$, are independent gaussian random variables centered at the value of the latent function $X_i(t)$ at time t_j with variance σ^2. Thus each observation has distribution

$$f(Y_i(t_j) \mid X_i(t_j), \sigma^2) = N(X_i(t_j), \sigma^2) \text{ for } i = 1, \ldots, N \quad j = 1, \ldots, p$$

which results in the joint distribution of all observations

$$f(\mathbf{Y} \mid \mathbf{X}, \sigma^2) = \prod_{i=1}^{N} N_p(\mathbf{X}_i, \sigma^2 I)$$

where $\mathbf{Y}$ is the matrix such that the observation for function $X_i(t)$ at time point t_j is in the ith row and the jth column, $\mathbf{X}$ is the matrix of the corresponding means for each entry in $\mathbf{Y}$, and $\mathbf{X}_i = (X_i(t_1), \ldots, X_i(t_p))'$, the vector of evaluations of the functions $X_i(t)$ at time points $\mathbf{t} = (t_1, \ldots, t_p)'$.

The following priors are assumed

$$\mathbf{X}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}_X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_X) \text{ for } i = 1, \ldots, N$$

$$\boldsymbol{\mu} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_\mu)$$

$$\boldsymbol{\Sigma}_X \sim IW(\mathbf{P}_X, \delta)$$

$$\sigma^2 \sim IG(a, b)$$

$$\eta_1 \sim IG(a, b) \quad \eta_2 \sim G(c, d)$$

$$\lambda_1 \sim IG(a, b) \quad \lambda_2 \sim G(c, d)$$

where a, b, c, and d are fixed hyperparameters and $\mathbf{\Sigma}_\mu$ and $\mathbf{P}_X$ are hyperparmeters that include smoothing information from the penalty matrix.

In the priors above, the roughness penalties for the latent and mean functions are specifically defined as (where $\mathbf{P}_1$ and $\mathbf{P}_2$ are defined in (4)):

$$\mathbf{P}_X = \eta_1^{-1}\mathbf{P}_1 + \lambda_1^{-1}\mathbf{P}_2 \quad \text{and} \quad \mathbf{\Sigma}_\mu = \eta_2^{-1}\mathbf{P}_1 + \lambda_2^{-1}\mathbf{P}_2$$

Using these assumptions, the following full conditional distributions are derived to run a MCMC Gibbs sampler, for $i = 1 \dots N$,

$$\mathbf{X}_i \mid rest \sim N_p((\sigma^{-2}I + \mathbf{\Sigma}_X^{-1})^{-1}(\sigma^{-2}Y_i + \mathbf{\Sigma}_X^{-1}\mu),(\sigma^{-2}I + \mathbf{\Sigma}_X^{-1})^{-1})$$

$$\boldsymbol{\mu} \mid rest \sim N_p((\mathbf{\Sigma}_\mu^{-1} + N\mathbf{\Sigma}_X^{-1})^{-1}(\mathbf{\Sigma}_X^{-1}\sum_{i=1}^N \mathbf{X}_i),(\mathbf{\Sigma}_\mu^{-1} + N\mathbf{\Sigma}_X^{-1})^{-1})$$

$$\mathbf{\Sigma}_X \mid rest \sim IW(\mathbf{P}_X + \sum_{i=1}^N(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})', \delta + N)$$

$$\sigma^2 \mid rest \sim IG(a + Np/2, b + 1/2\sum_{i=1}^N(Y_i - X_i)'(Y_i - X_i))$$

$$\eta_1 \mid rest \sim IG(a + p + 1, b + tr((\mathbf{P}_1\boldsymbol{\Sigma}_X^{-1})/2)$$

$$\lambda_1 \mid rest \sim IG(a + (p-2)(p+1)/2, b + tr((\mathbf{P}_2\boldsymbol{\Sigma}_X^{-1})/2)$$

$$\eta_2 \mid rest \sim G(c + 1, d + \boldsymbol{\mu}'\mathbf{P}_1^{-}\boldsymbol{\mu}/2)$$

$$\lambda_2 \mid rest \sim G(c + (p-2)/2, d + \boldsymbol{\mu}'\mathbf{P}_2^{-}\boldsymbol{\mu}/2)$$

A.2.2 Functional Regression

The model above can easily be extended to the framework of a functional linear regression model. With a scalar response, z_i and functional predictor $X_i(t)$, we are interested in finding a function $\beta(t)$ such that

$$z_i = \alpha + \int \beta(t)X_i(t)dt + \epsilon_i,\ i = 1, \ldots, N$$

$$\epsilon_i \sim N(0, \tau^2)$$

Again, the underlying assumption is that we observe noisy finite dimensional observations of the predictor $X_i(t)$ such that the distribution of the observations, $Y_i(t_j)$ is

$$f(Y_i(t_j) \mid X_i(t_j), \sigma^2) \sim N(X_i(t_j), \sigma^2) \text{ for } i = 1, \ldots, N \quad j = 1, \ldots, p$$

Using a finite approximation of the predictor, $X_i(t)$, let $\mathbf{X}$ equal the N x $(p + 1)$ matrix of predictors, $i = 1, \ldots, N$ where the first column is a column of ones and columns 2 through $j + 1$ consist of evaluations of $X_i(t)$ at t_j, $j = 1, \ldots, p$. In accordance, we will consider the $(p + 1)$ x 1 vector $\boldsymbol{\beta}$ such that $\boldsymbol{\beta} = (\alpha, \beta(t_1), \beta(t_2), \ldots, \beta(t_j))'$, a finite approximation of the functional regression coefficient such that $\alpha + \int \beta(t) X_i(t) dt \approx \mathbf{X}[i,]\boldsymbol{\beta}$, for each $i = 1, \ldots, N$. Under these assumptions the joint distribution of the independent observations, $\mathbf{z} = (z_1, \ldots, z_N)'$ and the N x p matrix $\mathbf{Y}$ is

$$f(\mathbf{z}, \mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}, \tau^2, \sigma^2) = N_N(\mathbf{X}\boldsymbol{\beta}, \tau^2 \mathbf{I}_N) \prod_{i=1}^{N} N_p(\mathbf{X}_i, \sigma^2 \mathbf{I}_p)$$

The following priors are assumed

$$\mathbf{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_X) \text{ for } i = 1, \ldots, N$$

$$\boldsymbol{\mu} \sim N_p(0, \boldsymbol{\Sigma}_\mu)$$

$$\boldsymbol{\Sigma}_X \sim IW(\mathbf{P}_X, \delta)$$

$$\boldsymbol{\beta} \sim N_{p+1}(\mathbf{0}, \boldsymbol{\Sigma}_\beta), \quad \boldsymbol{\Sigma}_\beta = \begin{pmatrix} \xi^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_\beta \end{pmatrix}$$

$$\sigma^2 \sim IG(a, b)$$

$$\tau^2 \sim IG(a, b)$$

$$\eta_1 \sim IG(a, b) \quad \eta_2 \sim G(c, d)$$

$$\lambda_1 \sim IG(a, b) \quad \lambda_2 \sim G(c, d) \quad \lambda_3 \sim G(c, d)$$

The hyperparameters ξ^2, a, b, c, and d are fixed and $\mathbf{\Sigma}_\mu$, $\mathbf{P}_X$ and $\mathbf{P}_\beta$ are hyperparmeters that include smoothing information from the penalty matrix. $\mathbf{\Sigma}_\mu$ and $\mathbf{P}_X$ are as defined in the section on estimating latent functions. Here, we have assumed for simplicity that $\mathbf{P}_\beta^{-1} = \lambda_3 \mathbf{\Sigma}_\mu^{-1}$. However, if a separate smoothing parameter for each element of the penalty matrix for the prior on $\boldsymbol{\beta}$ is desired, the additional smoothing parameter can easily be incorporated into this model.

For $i = 1, \ldots, N$, define

$$\boldsymbol{\beta}_X = \boldsymbol{\beta}[2 : (p + 1)] \text{ and } \mathbf{\Sigma}_{X_i|rest} = (\tau^{-2}\boldsymbol{\beta}_X\boldsymbol{\beta}_X' + \sigma^{-2}\mathbf{I}_p + \mathbf{\Sigma}_X^{-1})^{-1}$$

Then the full conditional distributions for the Gibbs Sampler are:

$$\mathbf{X}_i \mid rest \sim N_p(\mathbf{\Sigma}_{X_i|rest}(\tau^{-2}(z_i - \alpha)\boldsymbol{\beta}_X + \sigma^{-2}\mathbf{Y}_i + \mathbf{\Sigma}_X^{-1}\boldsymbol{\mu}), \mathbf{\Sigma}_{X_i|rest})$$

$$\mu \mid rest \sim N_p((N\boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Sigma}_\mu^{-1})^{-1}(\boldsymbol{\Sigma}_X^{-1} \sum_{i=1}^{N} \mathbf{X}_i), (N\boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Sigma}_\mu^{-1})^{-1})$$

$$\boldsymbol{\Sigma}_X \mid rest \sim IW(\mathbf{P}_X + \sum_{i=1}^{N}(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})', \delta + N)$$

$$\boldsymbol{\beta} \mid rest \sim N_p(\tau^{-2}(\tau^{-2}\mathbf{X}'\mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1})^{-1}\mathbf{X}'\mathbf{z}, (\tau^{-2}\mathbf{X}'\mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1})^{-1})$$

$$\sigma^2 \mid rest \sim IG(a + Np/2, b + 1/2 \sum_{i=1}^{N}(Y_i - X_i)'(Y_i - X_i))$$

$$\tau^2 \mid rest \sim IG(a + N/2, b + 1/2(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{z} - \mathbf{X}\boldsymbol{\beta}))$$

$$\eta_1 \mid rest \sim IG(a + p + 1, b + tr((\mathbf{P}_1\boldsymbol{\Sigma}_X^{-1})/2)$$

$$\lambda_1 \mid rest \sim IG(a + (p-2)(p+1)/2, b + tr((\mathbf{P}_2\boldsymbol{\Sigma}_X^{-1})/2)$$

$$\eta_2 \mid rest \sim G(c + 2, d + 1/2(\boldsymbol{\mu}'\mathbf{P}_1^{-}\boldsymbol{\mu} + \lambda_3\boldsymbol{\beta}_X'\mathbf{P}_1^{-}\boldsymbol{\beta}_X))$$

$$\lambda_2 \mid rest \sim G(c + p - 2, d + 1/2(\boldsymbol{\mu}'\mathbf{P}_2^{-}\boldsymbol{\mu} + \lambda_3\boldsymbol{\beta}_X'\mathbf{P}_2^{-}\boldsymbol{\beta}_X))$$

$$\lambda_3 \mid rest \sim G(c + p/2, d + 1/2\boldsymbol{\beta}_X'\boldsymbol{\Sigma}_\mu^{-1}\boldsymbol{\beta}_X)$$

A.2.3 Incorporating Missing Data

Assume the observed time points of observation i are $\mathbf{Y}_{io}$ and the missing data for observation i are denoted $\mathbf{Y}_{iu}$. Then, the joint distribution of the observed data is

$$f(\{\mathbf{Y}_{io} \mid i = 1, \ldots, N\} \mid \{\mathbf{X}_{io} \mid i = 1, \ldots, N\}, \sigma^2) = \prod_{i=1}^{N} N_{r_i}(\mathbf{X}_{io}, \sigma^2 \mathbf{I}_{r_i})$$

or in the case of functional linear regression

$$f(\mathbf{z}, \{\mathbf{Y}_{io} \mid i = 1 \ldots N\} \mid \mathbf{X}, \boldsymbol{\beta}, \tau^2, \sigma^2) = N_N(\mathbf{X}\boldsymbol{\beta}, \tau^2 \mathbf{I}_N) \prod_{i=1}^{N} N_{r_i}(\mathbf{X}_{io}, \sigma^2 \mathbf{I}_{r_i})$$

where r_i is the length of observed data for sample i.

Now in addition to the priors for the complete data smoothing problem outlined in the previous sections, the following prior for $\mathbf{Y}_{iu}$, $i = 1, \ldots, N$ is defined

$$\mathbf{Y}_{iu} \mid \mathbf{X}_{iu}, \sigma^2 \sim N_{p-r_i}(\mathbf{X}_{iu}, \sigma^2 \mathbf{I}_{p-r_i})$$

As $\mathbf{Y}_{iu}$ and $\mathbf{Y}_{io}$ are independent given $\mathbf{X}_i$ and σ^2 and no other distributions are dependent on $\mathbf{Y}_{iu}$, the full conditional distribution for $\mathbf{Y}_{iu}$ in this case takes the same form as the prior.

In any iteration, once the sample of $\mathbf{Y}_{iu}$ is drawn, the full vector of obser-

vations, $\mathbf{Y}_i$, can be reassembled by appropriately combining the data from the draw and observations, $\mathbf{Y}_{io}$. Now that $\mathbf{Y}_i$ is "known", the Gibbs sampler can proceed as if the vector, $\mathbf{Y}_i$, was observed in its entirety.

A.3 Figure: Credible Bands

As described in Section 2.4.2, ten sets of credible bands for the first eigenfunction and the mean function are used to empirically assess coverage properties in these models. Figure 2.4 contains plots of credible bands for five of the original ten subsets for both the first eigenfunction and the mean function. Figure A.1 includes similar plots for the remaining five subsets.

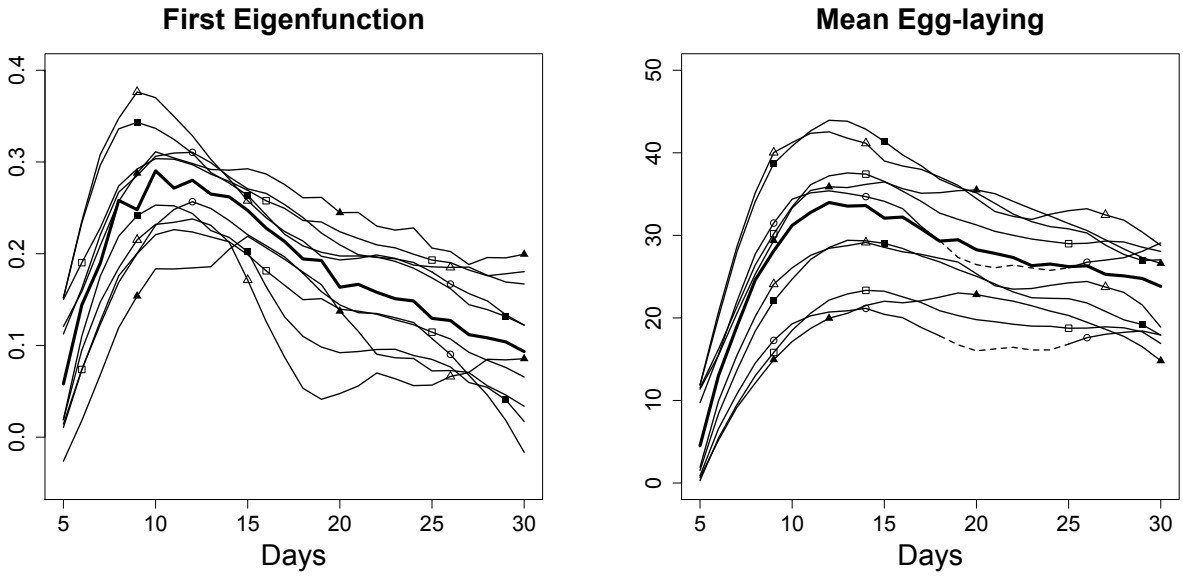

Figure A.1: Estimated credible interval coverage of the first eigenfunction (left) and the mean function (right) for the remaining five subsets. The thick lines in each plot is the first eigenfunction or mean function determined from the full data set of 534 medflies. Plotted with the population means for the first eigenfunction and the mean function respectively are 95% point wise credible bands for the corresponding function determined from each of the five remaining subsets of the original data, where the upper and lower credible bands for a particular subset are designated by matching symbols. The dashed lines highlight portions of time where a credible interval that does not contain the population mean.

APPENDIX B

APPENDIX TO CHAPTER 3

Below, in detail, are the specifications for the hierarchical Bayesian registration model discussed in Chapter 3. The first section includes the basic model for functional data registration also found in Section 3.1. Section B.2 describes the MCMC sampling scheme for this model.

B.1 Functional Registration

As discussed in Section 3.1, the initial assumption of this model is that we are interested in registering functional data, $X_i(t), i = 1, \ldots, N$, where these data are either observed directly over a set of time points, $\mathbf{t} = (t_1 \ldots t_p)'$, or are estimated from noisy observations, $\mathbf{Y}_i = (Y_i(t_1), \ldots, Y_i(t_p))'$. We assume a Gaussian noise process such that for each observation $\mathbf{Y}_i$,

$$f(Y_i(t_j) \mid X_i(t_j), \sigma^2) = N(X_i(t_j), \sigma^2) \text{ for } i = 1, \ldots, N \quad j = 1, \ldots, p$$

which results in the joint distribution of all observations

$$f(\mathbf{Y} \mid \mathbf{X}, \sigma^2) = \prod_{i=1}^{N} N_p(\mathbf{X}_i, \sigma^2 I)$$

where $\mathbf{Y}$ is the matrix such that the observation for function $X_i(t)$ at time

point t_j is in the ith row and the jth column, $\mathbf{X}$ is the matrix of the corresponding means for each entry in $\mathbf{Y}$, and $\mathbf{X}_i = (X_i(t_1), \ldots, X_i(t_p))'$, the vector of evaluations of the functions $X_i(t)$ at time points $\mathbf{t} = (t_1, \ldots, t_p)'$.

When the observations are observed noisily, the registered functions and noise variance are characterized by the following prior distributions:

$$\begin{aligned} \mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}, \gamma_R, \lambda_X &\sim N_p(z_{0i}\mathbf{1} + z_{1i}\mathbf{f}, \gamma_R^{-1}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_X) \quad i = 1 \ldots N \\ \boldsymbol{\Sigma}_X &= \eta_X^{-1}\mathbf{P}_1 + \lambda_X^{-1}\mathbf{P}_2 \\ \sigma_y^2 &\sim IG(a, b) \end{aligned} \tag{B.1}$$

However, If each function $X_i(t)$ is observed directly over $\mathbf{t}$, (B.1) assumes the roll of the distribution of the observed data and the covariance matrix, $\boldsymbol{\Sigma}_X$, designed to penalize roughness in the unregistered functions is excluded. This results in the following data distribution.

$$\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}, \gamma_R, \lambda_X \sim N_p(z_{0i}\mathbf{1} + z_{1i}\mathbf{f}, \gamma_R^{-1}\boldsymbol{\Sigma}) \quad i = 1 \ldots N \tag{B.2}$$

In both scenarios, we assume the following additional priors

$$
\begin{aligned}
\mathbf{h}_i(t_j) &= t_1 + \sum_{k=2}^{j}(t_k - t_{k-1})e^{w_i(t_k)} \quad i = 1 \ldots N \quad j = 1 \ldots p \\
\mathbf{w}_i &\propto N_{p-1}(\mathbf{0}, \gamma_w^{-1}\boldsymbol{\Sigma} + \lambda_w^{-1}\mathbf{P}_w)\mathbb{1}\{t_1 + \sum_{k=2}^{p}(t_k - t_{k-1})e^{w_i(t_k)} = t_p\} \\
&\quad i = 1 \ldots N \\
\mathbf{f} \mid \eta_f, \lambda_f &\sim N_p(0, \boldsymbol{\Sigma}_f) \\
\boldsymbol{\Sigma}_f &= \eta_f^{-1}\mathbf{P}_1 + \lambda_f^{-1}\mathbf{P}_2 \\
z_{0i} \mid \sigma_{z0}^2 &\sim N(0, \sigma_{z0}^2) \quad i = 1 \ldots (N-1) \quad z_{0N} = -\sum_{i=1}^{N-1} z_{0i} \\
\sigma_{z0}^2 &\sim IG(a, b) \\
z_{1i} \mid \sigma_{z1}^2 &\sim N(1, \sigma_{z1}^2) \quad i = 1 \ldots N \\
\sigma_{z1}^2 &\sim IG(a, b)] \\
\eta_f &\sim G(c, d) \\
\lambda_f &\sim G(c, d)
\end{aligned}
$$

where a, b, c, and d are fixed hyperparameters.

$\boldsymbol{\Sigma}$ is a fixed matrix designed to penalize variation in any direction from the corresponding mean of the distribution in which it is utilized. It is composed of two matrices, $\mathbf{P}_1$ and $\mathbf{P}_2$, such that $\boldsymbol{\Sigma} = \mathbf{P}_1 + \mathbf{P}_2$. $\mathbf{P}_1$ penalizes variation from the mean in constant and linear directions, and $\mathbf{P}_2$ penalizes variation from the mean in directions of curvature. For the distribution on the registered functions, $\boldsymbol{\Sigma}$ penalizes variation from a vertical shift and scaling of the target function. In the distribution of the base functions, $\boldsymbol{\Sigma}$ penalizes variation from the identity warping. The fixed parameters γ_R and γ_w determine the degree of these penalties for the registered functions and the base functions, respectively.

$\mathbf{P}_2$ is also used to penalize curvature in the registered functions, base functions, and the target function with associated smoothing parameters λ_X, λ_w, and λ_f. The exact definition of of $\mathbf{P}_1$ and $\mathbf{P}_2$ is found in equation (2.3). Also note, in the prior on the base functions (3.6), we are allowing $\mathbf{P}_w$ to be more generally interpreted as a penalty on severe deformations of the unregistered functions. Here, $\mathbf{P}_w$ may be either a penalty on the squared second derivative of the base functions (as above) or a penalty on the squared first derivative of the base functions.

B.2 MCMC Sampling

Using these assumptions, the following full conditional distributions are derived to run a MCMC sampler. Note, this list will not include a full conditional for the base functions or registered functions as their priors are not conjugate. Instead, the base and registered functions are sampled via a Metropolis step.

$$
\begin{aligned}
\mathbf{X}_i \mid rest &\sim N_p((\sigma_y^{-2}\mathbf{I}_p + \mathbf{\Sigma}_X^{-1})^{-1}(\sigma_y^{-2}\mathbf{I}_p\mathbf{Y}_i + \mathbf{\Sigma}_X^{-1}(z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}(\mathbf{h_i^{-1}})), (\sigma_y^{-2}\mathbf{I}_p + \mathbf{\Sigma}_X^{-1})^{-1}) \\
\mathbf{f} \mid rest &\sim N_p(\boldsymbol{\mu}_{\mathbf{f}|rest}, \mathbf{\Sigma}_{\mathbf{f}|rest}) \\
\mathbf{\Sigma}_{\mathbf{f}|rest} &= (\sum_{i=1}^{N} z_{1i}^2(\gamma_R^{-1}\mathbf{\Sigma} + \mathbf{\Sigma}_X)^{-1} + \mathbf{\Sigma}_f^{-1})^{-1} \\
\boldsymbol{\mu}_{\mathbf{f}|rest} &= \mathbf{\Sigma}_{\mathbf{f}|rest}((\gamma_R^{-1}\mathbf{\Sigma} + \mathbf{\Sigma}_X)^{-1} \sum_{i=1}^{N} z_{1i}(\mathbf{X_i}(\mathbf{h_i}) - z_{0i}\mathbf{1}_p)) \\
\sigma_Y^2 \mid rest &\sim IG(a + Np/2, b + 1/2 \sum_{i=1}^{N}(\mathbf{Y}_i - \mathbf{X}_i)'(\mathbf{Y}_i - \mathbf{X}_i)) \\
\eta_X \mid rest &\sim G(c + N, d + 1/2 \sum_{i=1}^{N} tr((\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}))(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}))'\mathbf{P}_1^-) \\
\lambda_X \mid rest &\sim G(c + N, d + 1/2 \sum_{i=1}^{N} tr((\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}))(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}))'\mathbf{P}_2^-) \\
z_{0i} \mid rest &\sim N(\mu_{z_{0i}|rest}, \sigma^2_{z_{0i}|rest}) \\
\sigma^2_{z_{0i}|rest} &= (\sigma_{z_0}^{-2} + 2 * \mathbf{1}_p'(\lambda_R^{-1}\mathbf{\Sigma} + \mathbf{\Sigma}_X)^{-1}\mathbf{1}_p)^{-1} \\
\mu_{z_{0i}|rest} &= \sigma^2_{z_{0i}|rest}(\mathbf{X_i}(\mathbf{h_i}) - \mathbf{X_N}(\mathbf{h_N}) + (z_{1N} - z_{1i})\mathbf{f} - \sum_{j=1}^{N-1} z_{0j}\mathbb{1}\{j \neq i\}\mathbf{1}_p)'(\gamma_R^{-1}\mathbf{\Sigma} + \mathbf{\Sigma}_X)^{-1}\mathbf{1}_p) \\
\sigma_{z_0}^2 \mid rest &\sim IG(a + (N-1)/2, b + 1/2 \sum_{i=1}^{N-1} z_{0i}^2) \\
z_{1i} \mid rest &\sim N(\mu_{z_{1i}|rest}, \sigma^2_{z_{1i}|rest}) \\
\sigma^2_{z_{1i}|rest} &= (\sigma_{z_1}^{-2} + \mathbf{f}'(\gamma_R^{-1}\mathbf{\Sigma} + \mathbf{\Sigma}_X)^{-1}\mathbf{f})^{-1} \\
\mu_{z_{1i}|rest} &= \sigma^2_{z_{1i}|rest}((\mathbf{X_i}(\mathbf{h_i}) - z_{0i}\mathbf{1}_p)'(\lambda_R^{-1}\mathbf{\Sigma} + \mathbf{\Sigma}_X)^{-1}\mathbf{f}) \\
\sigma_{z_1}^2 \mid rest &\sim IG(a + N/2, b + 1/2 \sum_{i=1}^{N} z_{1i}^2) \\
\eta_f \mid rest &\sim G(c + 1, d + (\mathbf{f}'\mathbf{P}_1^-\mathbf{f})/2) \\
\lambda_f \mid rest &\sim G(c + (p-2)/2, d + (\mathbf{f}'\mathbf{P}_2^-\mathbf{f})/2)
\end{aligned}
$$

B.3 Adapted Variational Bayes

Below are the approximate posterior distributions for the adapted variational Bayes estimation procedure outlined in Section 3.2.1. For a more thorough discussion and illustration of how the optimal q distributions are derived see Goldsmith et. al. [16].

As the variational Bayes procedure described in Section 3.2.1 requires conditionally conjugate distributions for all parameters except for the base functions, we will use the alternate expression for the smoothing piece of the distribution on the registered functions found in (3.14) to provide the means to apply adapted variational Bayes for the model where data is recorded with noise. This allows us to approximate the appropriate q distributions for $\mathbf{X}_i$, η_X, and λ_X. All other q distributions are determined in the usual way. For details on the approximated q distributions, see Section 3.2.1. Based on the full conditional distribution for above the following approximate posterior distributions are updated in

each iteration.

$$
\begin{aligned}
q(\mathbf{X}_i) &\sim N_p(\boldsymbol{\mu}_{q(\mathbf{X}_i)}, \boldsymbol{\Sigma}_{q(\mathbf{X}_i)}) \\
q(\mathbf{f}) &\sim N_p(\boldsymbol{\mu}_{q(\mathbf{f})}, \boldsymbol{\Sigma}_{q(\mathbf{f})}) \\
q(\sigma_Y^2) &\sim IG(a_{q(\sigma_Y^2)}, b_{q(\sigma_Y^2)}) \\
q(\eta_X) &\sim G(c_{q(\eta_X)}, d_{q(\eta_X)}) \\
q(\lambda_X) &\sim G(c_{q(\lambda_X)}, d_{q(\lambda_X)}) \\
q(z_{0i}) &\sim N(\mu_{q(z_{0i})}, \sigma^2_{q(z_{0i})}) \\
q(\sigma_{z_0}^2) &\sim IG(a_{q(\sigma_{z_0}^2)}, b_{q(\sigma_{z_0}^2)}) \\
q(z_{1i}) &\sim N(\mu_{q(z_{1i})}, \sigma^2_{q(z_{1i})}) \\
q(\sigma_{z_1}^2) &\sim IG(a_{q(\sigma_{z_1}^2)}, b_{q(\sigma_{z_1}^2)}) \\
q(\eta_f) &\sim G(c_{q(\eta_f)}, d_{q(\eta_f)}) \\
q(\lambda_f) &\sim G(c_{q(\lambda_f)}, d_{q(\lambda_f)})
\end{aligned}
$$

If the observations are recorded without noise, i.e. we have observations $\mathbf{X}_i$, $i = 1, \ldots, N$ as described in (B.2). The approximate posterior distribution of all parameters except the base functions is

$$
q(\boldsymbol{\theta}) = \prod_{k=(N+1)}^{d} q_k(\boldsymbol{\theta}_k) = q(\mathbf{f}) \prod_{i=1}^{(N-1)} q(z_{0i}) q(\sigma_{z_0}^2) \prod_{i=1}^{N} q(z_{1i}) q(\sigma_{z_1}^2) q(\eta_f) q(\lambda_f) \tag{B.3}
$$

If the observations have been recorded with noise, the additional required q densities are

$$
q(\mathbf{X}_i) \text{ for } i = 1, \ldots, N, q(\sigma_y^2), q(\eta_X), \text{ and } q(\lambda_X) \tag{B.4}
$$

As the q densities are all of known distributional forms, updating these densities is equivalent to updating their parameters. First, assuming the data are

recorded without noise, for each iteration, the following parameters are updated for the q densities found in the first part of this section. These updates are listed in an order that allows the convergence criterion to be calculated. Further details on the convergence criterion can be found in the next section.

$$\boldsymbol{\Sigma}_{q(\mathbf{f})} = \Big[\sum_{i=1}^{N} (\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})}) \gamma_R \boldsymbol{\Sigma}^{-1} + \mu_{q(\eta_\mathbf{f})} \mathbf{P}_1^- + \mu_{q(\lambda_\mathbf{f})} \mathbf{P}_2^- \Big]^{-1}$$

$$\boldsymbol{\mu}_{q(\mathbf{f})} = \boldsymbol{\Sigma}_{q(\mathbf{f})} \gamma_R \boldsymbol{\Sigma}^{-1} \Big[\sum_{i=1}^{N} \mu_{q(z_{1i})} (\mathbf{X}_i(\mathbf{h}_i) - \mu_{q(z_{0i})} \mathbf{1}_p) \Big]$$

$$\sigma^2_{q(z_{0i})} = (\mu_{q(\sigma_{z_0}^{-2})} + 2\mathbf{1}_p' \gamma_R \boldsymbol{\Sigma}^{-1} \mathbf{1}_p)^{-1}$$

$$\mu_{q(z_{0i})} = \sigma^2_{q(z_{0i})} (\mathbf{X}_i(\mathbf{h}_i) - \mathbf{X}_N(\mathbf{h}_N) + (\mu_{q(z_{1N})} - \mu_{q(z_{1i})}) \mu_{q(\mathbf{f})} - \sum_{j=1}^{N-1} \mu_{q(z_{0j})} \mathbb{1}\{i \neq j\} \mathbf{1}_p)$$

$$\sigma^2_{q(z_{1i})} = (\mu_{q(\sigma_{z_1}^{-2})} + tr((\boldsymbol{\Sigma}_{q(\mathbf{f})} + \mu_{q(\mathbf{f})} \mu'_{q(\mathbf{f})}) \gamma_R \boldsymbol{\Sigma}^{-1}))^{-1}$$

$$\mu_{q(z_{1i})} = \sigma^2_{q(z_{1i})} (\mu_{q(\sigma_{z_1}^{-2})} + \mu'_{q(\mathbf{f})} \gamma_R \boldsymbol{\Sigma}^{-1} (\mathbf{X}_i(\mathbf{h}_i) - \mu_{q(z_{0i})} \mathbf{1}_p))$$

$$d_{q(\eta_\mathbf{f})} = d + 1/2 * tr(\mathbf{P}_1^- (\Sigma_{q(\mathbf{f}_1)} + \mu_{q(\mathbf{f}_1)} \mu'_{q(\mathbf{f}_1)}))$$

$$d_{q(\lambda_\mathbf{f})} = d + 1/2 * tr(\mathbf{P}_2^- (\Sigma_{q(\mathbf{f}_1)} + \mu_{q(\mathbf{f}_1)} \mu'_{q(\mathbf{f}_1)}))$$

$$b_{q(\sigma^2_{z_0})} = b + 1/2 \sum_{i=1}^{N-1} (\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})})$$

$$b_{q(\sigma^2_{z_1})} = b + 1/2 \sum_{i=1}^{N} (\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})$$

For the model where observations are recorded with noise, in all of the updates above $\mathbf{X}_i(\mathbf{h}_i)$ and $\mathbf{X}_N(\mathbf{h}_N)$ are replaced by $\boldsymbol{\mu}_{q(\mathbf{X}_i(\mathbf{h}_i))}$ and $\boldsymbol{\mu}_{q(\mathbf{X}_N(\mathbf{h}_N))}$, respectively. Additionally, $\gamma_R \boldsymbol{\Sigma}^{-1}$ is replaced by $(\gamma_R^{-1} \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_X)^{-1}$. For each i, $\boldsymbol{\mu}_{q(\mathbf{X}_i(\mathbf{h}_i))}$ is determined by using the update for the mean of the q distribution of the unregistered function, $\boldsymbol{\mu}_{q(\mathbf{X}_i)}$ below, and registering it using the current value of $\mathbf{h}_i$. In addition to these modified updates, the following additional updates necessary:

$$\boldsymbol{\Sigma}_{q(\mathbf{X}_i)} = (\mu_{q(\frac{1}{\sigma^2_Y})}\mathbf{I}_p + \mu_{q(\eta_X)}\mathbf{P}_1^- + \mu_{q(\lambda_X)}\mathbf{P}_2^-)^{-1}$$

$$\boldsymbol{\mu}_{q(\mathbf{X}_i)} = \boldsymbol{\Sigma}_{q(\mathbf{X}_i)}[\mu_{q(\frac{1}{\sigma^2_Y})}\mathbf{Y}_i + (\mu_{q(\eta_X)}\mathbf{P}_1^- + \mu_{q(\lambda_X)}\mathbf{P}_2^-)(\mu_{q(z_{0i})}\mathbf{1}_p + \mu_{q(z_{1i})}E_{(\theta_{-\mathbf{X}_i})}[\mathbf{f}(\mathbf{h}_i^{-1})])]$$

$$b_{q(\sigma^2_Y)} = b + \frac{1}{2}\sum_{i=1}^{N}\Big(\mathbf{Y}_i'\mathbf{Y}_i - 2\boldsymbol{\mu}'_{q(\mathbf{X}_i)}\mathbf{Y}_i + \sum_{j=1}^{p}\boldsymbol{\Sigma}_{q(\mathbf{X}_i)}[j,j] + \boldsymbol{\mu}_{q(\mathbf{X}_i)}[j]^2\Big)$$

$$\begin{aligned} d_{q(\eta_\mathbf{X})} = d + \frac{1}{2}tr\Big[\sum_{i=1}^{N}\Big(&\boldsymbol{\Sigma}_{q(\mathbf{X}_i)} + \boldsymbol{\mu}_{q(\mathbf{X}_i)}\boldsymbol{\mu}'_{q(\mathbf{X}_i)} - 2\boldsymbol{\mu}_{q(\mathbf{X}_i)}(\mu_{q(z_{0i})}\mathbf{1}_p + \mu_{q(z_{1i})}E_{(\theta_{-\eta_X})}[\mathbf{f}(\mathbf{h}_i^{-1})])' \\ &+ (\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})})\mathbf{1}_p\mathbf{1}'_p + 2\mu_{q(z_{0i})}\mu_{q(z_{1i})}\mathbf{1}_pE_{(\theta_{-\eta_X})}[\mathbf{f}(\mathbf{h}_i^{-1})]' \\ &+ (\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})E_{(\theta_{-\eta_X})}[\mathbf{f}(\mathbf{h}_i^{-1})\mathbf{f}(\mathbf{h}_i^{-1})']\Big)\mathbf{P}_1^-\Big] \end{aligned}$$

$$\begin{aligned} d_{q(\lambda_\mathbf{X})} = d + \frac{1}{2}tr\Big[\sum_{i=1}^{N}\Big(&\boldsymbol{\Sigma}_{q(\mathbf{X}_i)} + \boldsymbol{\mu}_{q(\mathbf{X}_i)}\boldsymbol{\mu}'_{q(\mathbf{X}_i)} - 2\boldsymbol{\mu}_{q(\mathbf{X}_i)}(\mu_{q(z_{0i})}\mathbf{1}_p + \mu_{q(z_{1i})}E_{(\theta_{-\lambda_X})}[\mathbf{f}(\mathbf{h}_i^{-1})])' \\ &+ (\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})})\mathbf{1}_p\mathbf{1}'_p + 2\mu_{q(z_{0i})}\mu_{q(z_{1i})}\mathbf{1}_pE_{(\theta_{-\lambda_X})}[\mathbf{f}(\mathbf{h}_i^{-1})]' \\ &+ (\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})E_{(\theta_{-\lambda_X})}[\mathbf{f}(\mathbf{h}_i^{-1})\mathbf{f}(\mathbf{h}_i^{-1})']\Big)\mathbf{P}_2^-\Big] \end{aligned}$$

Note, these updates contain terms that cannot be evaluated. For instance, $E_{(\theta_{-\eta_X})}[\mathbf{f}(\mathbf{h}_i^{-1})\mathbf{f}(\mathbf{h}_i^{-1})']$ cannot be determined because the approximate distribution of $\mathbf{f}(\mathbf{h}_i^{-1})$ is unknown. These terms can however be approximated. Section 3.5.2 provides details of the approximated values used for this analysis.

B.4 Convergence Criterion

When the functional observations, $\mathbf{X} = \mathbf{X}_i$, $i \ldots N$, are recorded without noise $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})]$ is monitored until the desired threshhold is met.

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ (f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}}) f(\boldsymbol{\theta}_{-\mathbf{w}})) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}}) + log\ f(\boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}})] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{f}) - log\ q(\mathbf{f})] \\
&\quad + \sum_{i=1}^{(N-1)} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(z_{0i}) - log\ q(z_{0i})] \\
&\quad + \sum_{i=1}^{(N)} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(z_{1i}) - log\ q(z_{1i})] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\sigma^2_{z_0}) - log\ q(\sigma^2_{z_0})] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\sigma^2_{z_1}) - log\ q(\sigma^2_{z_1})] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\eta_f) - log\ q(\eta_f)] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\lambda_f) - log\ q(\lambda_f)]
\end{aligned}
$$

Now looking at each piece individually,

$$
\begin{aligned}
& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}})] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\sum_{i=1}^{N}(log[(2\pi)^{-p/2} \mid \gamma_R^{-1}\boldsymbol{\Sigma} \mid^{-1/2}])\Big] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\sum_{i=1}^{N} -\frac{1}{2}[(\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}))'\gamma_R\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i(\mathbf{h}_i) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}))]\Big] \\
&= \sum_{i=1}^{N}(log[(2\pi)^{-p/2} \mid \gamma_R^{-1}\boldsymbol{\Sigma} \mid^{-1/2}]) \\
&\quad + \sum_{i=1}^{N} -\frac{1}{2}\Big[(\mathbf{X}_i(\mathbf{h}_i)'\gamma_R\boldsymbol{\Sigma}^{-1}\mathbf{X}_i(\mathbf{h}_i)) - \\
&\quad 2\mathbf{X}_i(\mathbf{h}_i)'\gamma_R\boldsymbol{\Sigma}^{-1}\mu_{q(z_{0i})}\mathbf{1}_p - 2\mathbf{X}_i(\mathbf{h}_i)'\gamma_R\boldsymbol{\Sigma}^{-1}\mu_{q(z_{1i})}\boldsymbol{\mu}_{q(\mathbf{f})} + \\
&\quad (\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})tr((\boldsymbol{\Sigma}_{q(\mathbf{f})} + \boldsymbol{\mu}_{q(\mathbf{f})}\boldsymbol{\mu}'_{q(\mathbf{f})})\gamma_R\boldsymbol{\Sigma}^{-1}) + \\
&\quad 2\mu_{q(z_{0i})}\mu_{q(z_{1i})}\mathbf{1}'_p\gamma_R\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{q(\mathbf{f})}\Big] - \\
&\quad \Big[\sum_{i=1}^{N-1}(\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})}) + \frac{1}{2}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\mu_{q(z_{0i})}\mu_{q(z_{0j})}\mathbb{1}\{j \neq i\}\Big]\mathbf{1}'_p\gamma_R\boldsymbol{\Sigma}^{-1}\mathbf{1}_p
\end{aligned}
$$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{f}) - log\ q(\mathbf{f})] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{p}{2}log\ 2\pi + \frac{1}{2}log\ \mid \eta_f\mathbf{P}_1^- + \lambda_f\mathbf{P}_2^- \mid\Big] - \\
&\quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}(tr[\mathbf{f}\mathbf{f}'(\eta_f\mathbf{P}_1^- + \lambda_f\mathbf{P}_2^-)]\Big] + \\
&\quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{p}{2}log\ 2\pi + \frac{1}{2}log\ \mid \boldsymbol{\Sigma}_{q(\mathbf{f})} \mid\Big] + \\
&\quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}tr(\mathbf{f}\mathbf{f}'\boldsymbol{\Sigma}^{-1}_{q(\mathbf{f})}) - \mathbf{f}'\boldsymbol{\Sigma}^{-1}_{q(\mathbf{f})}\boldsymbol{\mu}_{q(\mathbf{f})}\Big] + \\
&\quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}\boldsymbol{\mu}'_{q(\mathbf{f})}\boldsymbol{\Sigma}^{-1}_{q(\mathbf{f})}\boldsymbol{\mu}_{q(\mathbf{f})}\Big] \\
&= C + \frac{1}{2}E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[2log\ \eta_f] + \frac{1}{2}E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[(p-2)log\ \lambda_f] - \\
&\quad \frac{1}{2}tr\Big((\boldsymbol{\Sigma}_{q(\mathbf{f})} + \boldsymbol{\mu}_{q(\mathbf{f})}\boldsymbol{\mu}'_{q(\mathbf{f})})(\mu_{q(\eta_f)}\mathbf{P}_1^- + \mu_{q(\lambda_f)}\mathbf{P}_2^-)\Big) - \\
&\quad \frac{1}{2}log\ \mid \boldsymbol{\Sigma}^{-1}_{q(\mathbf{f})} \mid + \frac{p}{2}
\end{aligned}
$$

where C is a constant that does not change from one iteration to the next. For $\mathbf{z}_0 = (z_{01}, \ldots, z_{0(N-1)})'$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{z}_0) - log\ q(\mathbf{z}_0)] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{N-1}{2} log\ 2\pi - \frac{N-1}{2} log\ \sigma^2_{z_0} - \sum_{i=1}^{N-1} -\frac{1}{2\sigma^2_{z_0}} z^2_{0i} + \\
&\quad \frac{N-1}{2} log\ 2\pi + \frac{N-1}{2} log\ \sigma^2_{q(z_{0i})} + \\
&\quad \sum_{i=1}^{N-1} \frac{1}{2\sigma^2_{q(z_{0i})}} (z_{0i} - \mu_{q(z_{0i})})^2 \Big] \qquad \text{(B.5)} \\
&= \frac{N-1}{2} log\ \sigma^2_{q(z_{0i})} - E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{N-1}{2} log\ \sigma^2_{z_0}\Big] - \qquad \text{(B.6)} \\
&\quad \frac{1}{2}\mu_{q(\frac{1}{\sigma^2_{z_0}})}\Big(\sum_{i=1}^{N-1} (\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})})\Big) + \frac{N-1}{2}
\end{aligned}
$$

For $\mathbf{z}_1 = (z_{11}, \ldots, z_{1N})'$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{z}_1) - log\ q(\mathbf{z}_1)] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{N}{2} log\ 2\pi - \frac{N}{2} log\ \sigma^2_{z_1} - \sum_{i=1}^{N} -\frac{1}{2\sigma^2_{z_1}} z^2_{1i} + \\
&\quad \frac{N}{2} log\ 2\pi + \frac{N}{2} log\ \sigma^2_{q(z_1)} + \sum_{i=1}^{N} \frac{1}{2\sigma^2_{q(z_{1i})}} (z_{1i} - \mu_{q(z_{1i})})^2 \Big] \\
&= \frac{N}{2} log\ \sigma^2_{q(z_{1i})} - E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{N}{2} log\ \sigma^2_{z_1}\Big] - \\
&\quad \frac{1}{2}\mu_{q(\frac{1}{\sigma^2_{z_1}})}\Big(\sum_{i=1}^{N} (\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})\Big) + \frac{N}{2}
\end{aligned}
$$

For $\sigma^2_{z_0}$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\sigma^2_{z_0}) - log\ q(\sigma^2_{z_0})] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{b^a}{\Gamma(a)} - (a+1)log\ \sigma^2_{z_0} - b\frac{1}{\sigma^2_{z_0}} - \\
&\quad log\ \frac{b^{a_{q(\sigma^2_{z_0})}}_{q(\sigma^2_{z_0})}}{\Gamma(a_{q(\sigma^2_{z_0})})} + (a_{q(\sigma^2_{z_0})}+1)log\ \sigma^2_{z_0} + \\
&\quad b_{q(\sigma^2_{z_0})}\frac{1}{\sigma^2_{z_0}}\Big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[\ -(a+1)log\ \sigma^2_{z_0}] - b\mu_{q(\frac{1}{\sigma^2_{z_0}})} - log\ \frac{b^{a_{q(\sigma^2_{z_0})}}_{q(\sigma^2_{z_0})}}{\Gamma(a_{q(\sigma^2_{z_0})})} + \\
&\quad log\ \frac{b^a}{\Gamma(a)} + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[(a_{q(\sigma^2_{z_0})}+1)log\ \sigma^2_{z_0}] + b_{q(\sigma^2_{z_0})}\mu_{q(\frac{1}{\sigma^2_{z_0}})}
\end{aligned}
$$

For $\sigma^2_{z_1}$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\sigma^2_{z_1}) - log\ q(\sigma^2_{z_1})] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{b^a}{\Gamma(a)} - (a+1)log\ \sigma^2_{z_1} - b\frac{1}{\sigma^2_{z_1}} - \\
&\quad log\ \frac{b^{a_{q(\sigma^2_{z_1})}}_{q(\sigma^2_{z_1})}}{\Gamma(a_{q(\sigma^2_{z_1})})} + (a_{q(\sigma^2_{z_1})}+1)log\ \sigma^2_{z_1} + \\
&\quad b_{q(\sigma^2_{z_1})}\frac{1}{\sigma^2_{z_1}}\Big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[\ -(a+1)log\ \sigma^2_{z_1}] - b\mu_{q(\frac{1}{\sigma^2_{z_1}})} - log\ \frac{b^{a_{q(\sigma^2_{z_1})}}_{q(\sigma^2_{z_1})}}{\Gamma(a_{q(\sigma^2_{z_1})})} + \\
&\quad log\ \frac{b^a}{\Gamma(a)} + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[(a_{q(\sigma^2_{z_1})}+1)log\ \sigma^2_{z_1}] + b_{q(\sigma^2_{z_1})}\mu_{q(\frac{1}{\sigma^2_{z_1}})}
\end{aligned}
$$

For η_f

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\eta_f) - log\ q(\eta_f)] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{d^c}{\Gamma(c)} + (c-1)log\ \eta_f - d\eta_f - \\
&\quad log\ \frac{d_{q(\eta_f)}^{c_{q(\eta_f)}}}{\Gamma(c_{q(\eta_f)})} - c\ log\ \eta_f + d_{q(\eta_f)}\eta_f\Big] \\
&= log\ \frac{d^c}{\Gamma(c)} - log\ \frac{d_{q(\eta_f)}^{c_{q(\eta_X)}}}{\Gamma(c_{q(\eta_f)})} - E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ \eta_f] - d\mu_{q(\eta_f)} + \\
&\quad d_{q(\eta_f)}\mu_{q(\eta_f)}
\end{aligned}
$$

For λ_f

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\lambda_f) - log\ q(\lambda_f)] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{d^c}{\Gamma(c)} + (c-1)log\ \lambda_f - d\lambda_f - \\
&\quad log\ \frac{d_{q(\lambda_f)}^{c_{q(\lambda_f)}}}{\Gamma(c_{q(\lambda_f)})} - \Big(\frac{p-2}{2} + c - 1\Big)\ log\ \lambda_f + d_{q(\lambda_f)}\lambda_f\Big] \\
&= log\ \frac{d^c}{\Gamma(c)} - log\ \frac{d_{q(\lambda_f)}^{c_{q(\lambda_f)}}}{\Gamma(c_{q(\lambda_f)})} - \frac{p-2}{2}E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ \lambda_f] - d\mu_{q(\lambda_f)} + \\
&\quad d_{q(\lambda_f)}\mu_{q(\lambda_f)}
\end{aligned}
$$

The expression for $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})]$ can be simplified much further by combining terms that cancel out. However, in some cases the ability to cancel terms depends on the order of the updates. For instance, in the expression, $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\sigma_{z_0}^2) - log\ q(\sigma_{z_0}^2)]$, the terms $-b\mu_{q(\frac{1}{\sigma_{z_0}^2})}$ and $b_{q(\sigma_{z_0}^2)}\mu_{q(\frac{1}{\sigma_{z_0}^2})}$ cancel with $-\frac{1}{2}\mu_{q(\frac{1}{\sigma_{z_0}^2})}\Big(\sum_{i=1}^{N-1}(\sigma_{q(z_{0i})}^2 + \mu_{q(z_{0i})}^2)\Big)$ from $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{z}_0) - log\ q(\mathbf{z}_0)]$ as long as the parameters of $q(\mathbf{z}_0)$ are updated before $b_{q(\sigma_{z_0}^2)}$. For convenience, we have taken account the ordering necessary to compute the convergence criterion in the updates given above. Additionally, note all components in this expression that do not change from one iteration to the next can be ignored.

APPENDIX C

APPENDIX TO CHAPTER 4

Below, in detail, are the specifications for the hierarchical Bayesian registration model discussed in this paper. The first section includes the basic model for functional data registration also found in Section 4.1. Section C.2 describes the MCMC sampling scheme for this model.

C.1 Factor Analysis

As discussed in Section 4.1, the initial assumption of this model is that we are interested in registering and possibly clustering functional data, $X_i(t), i = 1, \ldots, N$. The registered functions, $X_i(h_i(t))$, $i = 1 \ldots N$, are assumed to be characterized almost completely by a linear combination of two factors, $f_1(t)$ and $f_2(t)$. Below

are the data and prior distributions used for this model.

$$
\begin{aligned}
\mathbf{X}_i(\mathbf{h}_i) \mid z_{0i}, z_{1i}, \mathbf{f}_1, z_{2i}, \mathbf{f}_2, \gamma_1, \gamma_2 &\sim N_p(z_{0i}\mathbf{1} + z_{1i}\mathbf{f}_1 + \frac{\gamma_2}{\gamma_1+\gamma_2} z_{2i}\mathbf{f}_2, (\gamma_1+\gamma_2)^{-1}\mathbf{\Sigma}) \quad i = 1 \ldots N \\
\mathbf{h}_i(t_j) &= t_1 + \sum_{k=2}^{j} (t_k - t_{k-1}) e^{w_i(t_k)} \quad i = 1 \ldots N \quad j = 1 \ldots p \\
\mathbf{w}_i \mid \gamma_w &\propto N_{p-1}(\mathbf{0}, \gamma_w^{-1}\mathbf{\Sigma} + \lambda_w^{-1}\mathbf{P}_2)\mathbb{1}\{t_1 + \sum_{k=2}^{p} (t_k - t_{k-1}) e^{w_i(t_k)} = t_p\} \quad i = 1 \ldots N \\
z_{0i} \mid \sigma^2_{z0} &\sim N(0, \sigma^2_{z0}) \quad i = 1 \ldots (N-1) \quad z_{0N} = -\sum_{i=1}^{N-1} z_{0i} \\
\sigma^2_{z0} &\sim IG(a, b) \\
z_{1i} \mid \sigma^2_{z1} &\sim N(1, \sigma^2_{z1}) \quad i = 1 \ldots N \\
\sigma^2_{z1} &\sim IG(a, b) \\
z_{2i} \mid \sigma^2_{z2} &\sim N(1, \sigma^2_{z2}) \quad i = 1 \ldots N \\
\sigma^2_{z2} &\sim IG(a, b) \\
\mathbf{f}_1 \mid \eta_f, \lambda_f &\sim N_p(0, \mathbf{\Sigma}_f) \\
\mathbf{f}_2 \mid \eta_f, \lambda_f &\sim N_p(0, \mathbf{\Sigma}_f) \\
\mathbf{\Sigma}_f &= \eta_f^{-1}\mathbf{P}_1 + \lambda_f^{-1}\mathbf{P}_2
\end{aligned}
$$

$\mathbf{\Sigma}$ is a fixed matrix designed to penalize variation in any direction from the corresponding mean of the distribution in which it is utilized. It is composed of two matrices, $\mathbf{P}_1$ and $\mathbf{P}_2$, such that $\mathbf{\Sigma} = \mathbf{P}_1 + \mathbf{P}_2$. $\mathbf{P}_1$ penalizes variation from the mean in constant and linear directions, and $\mathbf{P}_2$ penalizes variation from the mean in directions of curvature. For the distribution on the registered functions, $\mathbf{\Sigma}$ penalizes variation from a vertical shift and scaling of the target function. In the distribution of the base functions, $\mathbf{\Sigma}$ penalizes variation from the identity warping. The fixed parameters γ_R and γ_w determine the degree of these penalties for the registered functions and the base functions, respectively.

$\mathbf{P}_2$ is also used to penalize curvature in the registered functions, base functions, and the target function with associated smoothing parameters λ_X, λ_w, and λ_f. Further details of the construction of $\mathbf{P}_1$ and $\mathbf{P}_2$ are found in Earls and Hooker [9].

C.2 MCMC Sampling

Using these assumptions, the following full conditional distributions are derived to run a MCMC sampler. Note, this list will not include a full conditional for the base functions or registered functions as their priors are not conjugate. Instead, the base and registered functions are sampled via a Metropolis step.

$$
\begin{aligned}
\mathbf{f}_1 \mid rest &\sim N_p(\boldsymbol{\mu}_{\mathbf{f}_1|rest}, \boldsymbol{\Sigma}_{\mathbf{f}_1|rest}) \\
\boldsymbol{\Sigma}_{\mathbf{f}_1|rest} &= (\sum_{i=1}^{N} z_{1i}^2(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_f^{-1})^{-1} \\
\boldsymbol{\mu}_{\mathbf{f}_1|rest} &= \boldsymbol{\Sigma}_{\mathbf{f}_1|rest}\Big[(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N} z_{1i}\Big(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1} + \frac{\gamma_2}{\gamma_1+\gamma_2}z_{2i}\mathbf{f}_2)\Big)\Big] \\
\mathbf{f}_2 \mid rest &\sim N_p(\boldsymbol{\mu}_{\mathbf{f}_2|rest}, \boldsymbol{\Sigma}_{\mathbf{f}_2|rest}) \\
\boldsymbol{\Sigma}_{\mathbf{f}_2|rest} &= \Big(\sum_{i=1}^{N} z_{2i}^2\Big(\frac{\gamma_2^2}{\gamma_1+\gamma_2}\Big)\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_f^{-1}\Big)^{-1} \\
\boldsymbol{\mu}_{\mathbf{f}_2|rest} &= \boldsymbol{\Sigma}_{\mathbf{f}_2|rest}\Big[\gamma_2\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N} z_{2i}\Big(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1} + z_{1i}\mathbf{f}_1)\Big)\Big] \\
z_{0i} \mid rest &\sim N(\mu_{z_{0i}|rest}, \sigma^2_{z_{0i}|rest}) \\
\sigma^2_{z_{0i}|rest} &= (\sigma_{z_0}^{-2} + 2 * \mathbf{1}_p'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p)^{-1} \\
\mu_{z_{0i}|rest} &= \sigma^2_{z_{0i}|rest}\Big(\mathbf{X_i}(\mathbf{h_i}) - \mathbf{X_N}(\mathbf{h_N}) + (z_{1N}-z_{1i})\mathbf{f}_1 + \Big(\frac{\gamma_2}{\gamma_1+\gamma_2}\Big)(z_{2N}-z_{2i})\mathbf{f}_2 - \\
&\qquad \sum_{j=1}^{N-1} z_{0j}\mathbb{1}\{j\neq i\}\mathbf{1}_p\Big)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p \\
\sigma^2_{z_0} \mid rest &\sim IG(a + (N-1)/2, b + 1/2\sum_{i=1}^{N-1} z_{0i}^2) \\
z_{1i} \mid rest &\sim N(\mu_{z_{1i}|rest}, \sigma^2_{z_{1i}|rest}) \\
\sigma^2_{z_{1i}|rest} &= (\sigma_{z_1}^{-2} + \mathbf{f}_2'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{f}_2)^{-1} \\
\mu_{z_{2i}|rest} &= \sigma^2_{z_{2i}|rest}\Big(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1}_p + \frac{\gamma_2}{\gamma_1+\gamma_2}z_{2i}\mathbf{f}_2)\Big)'(\gamma_1+\gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{f}_1 \\
\sigma^2_{z_1} \mid rest &\sim IG(a + N/2, b + 1/2\sum_{i=1}^{N} z_{1i}^2) \\
z_{2i} \mid rest &\sim N(\mu_{z_{2i}|rest}, \sigma^2_{z_{2i}|rest}) \\
\sigma^2_{z_{2i}|rest} &= (\sigma_{z_2}^{-2} + \mathbf{f}_2'\frac{\gamma_2^2}{\gamma_1+\gamma_2}\boldsymbol{\Sigma}^{-1}\mathbf{f}_2)^{-1} \\
\mu_{z_{2i}|rest} &= \sigma^2_{z_{2i}|rest}\gamma_2\Big(\mathbf{X_i}(\mathbf{h_i}) - (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1)\Big)'\boldsymbol{\Sigma}^{-1}\mathbf{f}_2 \\
\sigma^2_{z_2} \mid rest &\sim IG(a + N/2, b + 1/2\sum_{i=1}^{N} z_{2i}^2) \\
\eta_f \mid rest &\sim G(c+2, d + \frac{1}{2}tr((\mathbf{f}_1\mathbf{f}_1' + \mathbf{f}_2\mathbf{f}_2')\mathbf{P}_1^-)) \\
\lambda_f \mid rest &\sim G(c+(p-2), d + \frac{1}{2}tr((\mathbf{f}_1\mathbf{f}_1' + \mathbf{f}_2\mathbf{f}_2')\mathbf{P}_2^-))
\end{aligned}
$$

C.3 Adapted Variational Bayes

After initializing all parameters, in each iteration, the adapted variational Bayes algorithm performs two steps. In the first step, the 'likelihood' as a function of the base functions is maximized. For this 'likelihood', all other parameters are fixed at their current values. The second step uses a traditional variational Bayes iterative scheme to update all other parameters. Specifically, assuming $\boldsymbol{\theta}_k = \mathbf{w}_k$, for $k = 1 \ldots N$, so that, $\boldsymbol{\theta} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N, \boldsymbol{\theta}_{N+1}, \ldots, \boldsymbol{\theta}_d\}$, the adapted variational Bayes algorithm is as follows:

1. Initialize $\boldsymbol{\theta}$
2. For each iteration, m, and each k, $k = 1 \ldots N$, update the estimate for $\mathbf{w}_k$ so that $\mathbf{w}_k^{(m)} = \sup_{\mathbf{w}_k} q_k(\mathbf{w}_k \mid \boldsymbol{\theta}_j^{(m-1)}, j = (N+1) \ldots d)$
3. For each iteration, m, and each k, $k = (N+1) \ldots d$, update q_k so that $q_k^{(m)} \propto exp[E_{(\boldsymbol{\theta}_{-k})}(log\ f(\boldsymbol{\theta}_k \mid rest)]$, where the expectation is taken with respect to the distributions $q_j^{(m-1)}(\boldsymbol{\theta}_j)$, $j = 1 \ldots d$, $j \neq k$
4. Repeat steps (2) and (3) until the desired convergence criterion is met

Below are the approximate posterior distributions, $q_k(\boldsymbol{\theta}_k)$, $k = (N+1), \ldots, d$, for the adapted variational Bayes estimation procedure described in Section 3.2.1. Note, the subscripts on the q distributions has been omitted. For a more thorough discussion and illustration of how the optimal q distributions are derived see Goldsmith et. al. [16].

$$
\begin{aligned}
q(\mathbf{f}_1) &\sim N_p(\boldsymbol{\mu}_{q(\mathbf{f}_1)}, \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}) \\
q(\mathbf{f}_2) &\sim N_p(\boldsymbol{\mu}_{q(\mathbf{f}_2)}, \boldsymbol{\Sigma}_{q(\mathbf{f}_2)}) \\
q(z_{0i}) &\sim N(\mu_{q(z_{0i})}, \sigma^2_{q(z_{0i})}) \\
q(\sigma^2_{z_0}) &\sim IG(a_{q(\sigma^2_{z_0})}, b_{q(\sigma^2_{z_0})}) \\
q(z_{1i}) &\sim N(\mu_{q(z_{1i})}, \sigma^2_{q(z_{1i})}) \\
q(\sigma^2_{z_1}) &\sim IG(a_{q(\sigma^2_{z_1})}, b_{q(\sigma^2_{z_1})}) \\
q(z_{2i}) &\sim N(\mu_{q(z_{2i})}, \sigma^2_{q(z_{2i})}) \\
q(\sigma^2_{z_2}) &\sim IG(a_{q(\sigma^2_{z_2})}, b_{q(\sigma^2_{z_2})}) \\
q(\eta_f) &\sim G(c_{q(\eta_f)}, d_{q(\eta_f)}) \\
q(\lambda_f) &\sim G(c_{q(\lambda_f)}, d_{q(\lambda_f)})
\end{aligned}
$$

The approximate joint posterior distribution of all parameters except the base functions is

$$
q(\boldsymbol{\theta}) = \prod_{k=(N+1)}^{d} q_k(\boldsymbol{\theta}_k) = q(\mathbf{f}_1)q(\mathbf{f}_2)q(\sigma^2_{z_0})q(\sigma^2_{z_1})q(\sigma^2_{z_2})q(\eta_f)q(\lambda_f) \prod_{i=1}^{(N-1)} q(z_{0i}) \prod_{i=1}^{N} q(z_{1i})q(z_{2i}) \quad \text{(C.1)}
$$

As the q densities are all of known distributional forms, updating these densities is equivalent to updating their parameters. For each iteration, the following parameters are updated for the q densities found in (C.1). These updates are listed in an order that allows the convergence criterion to be calculated. Further details on the convergence criterion can be found in the next section.

$$\boldsymbol{\Sigma}_{q(\mathbf{f}_1)} = \Big[\sum_{i=1}^{N} (\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1} + \mu_{q(\eta_\mathbf{f})}\mathbf{P}_1^- + \mu_{q(\lambda_\mathbf{f})}\mathbf{P}_2^- \Big]^{-1}$$

$$\boldsymbol{\mu}_{q(\mathbf{f}_1)} = \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}\Big[\sum_{i=1}^{N} \mu_{q(z_{1i})}(\mathbf{X}_i(\mathbf{h}_i) - (\mu_{q(z_{0i})}\mathbf{1}_p + \frac{\gamma_2}{\gamma_1 + \gamma_2}\mu_{q(z_{2i})}\boldsymbol{\mu}_{q(\mathbf{f}_2)}))\Big]$$

$$\boldsymbol{\Sigma}_{q(\mathbf{f}_2)} = \Big[\sum_{i=1}^{N} (\sigma^2_{q(z_{2i})} + \mu^2_{q(z_{2i})})\frac{\gamma_2^2}{\gamma_1 + \gamma_2}\boldsymbol{\Sigma}^{-1} + \mu_{q(\eta_\mathbf{f})}\mathbf{P}_1^- + \mu_{q(\lambda_\mathbf{f})}\mathbf{P}_2^- \Big]^{-1}$$

$$\boldsymbol{\mu}_{q(\mathbf{f}_2)} = \boldsymbol{\Sigma}_{q(\mathbf{f}_2)}\gamma_2\boldsymbol{\Sigma}^{-1}\Big[\sum_{i=1}^{N} \mu_{q(z_{2i})}(\mathbf{X}_i(\mathbf{h}_i) - (\mu_{q(z_{0i})}\mathbf{1}_p + \mu_{q(z_{1i})}\boldsymbol{\mu}_{q(\mathbf{f}_1)}))\Big]$$

$$\sigma^2_{q(z_{0i})} = (\mu_{q(\sigma^{-2}_{z_0})} + \mathbf{1}'_p(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p)^{-1}$$

$$\mu_{q(z_{0i})} = \sigma^2_{q(z_{0i})}(\mathbf{X}_i(\mathbf{h}_i) - \mathbf{X}_N(\mathbf{h}_N) + (\mu_{q(z_{1N})} - \mu_{q(z_{1i})})\mu_{q(\mathbf{f}_1)} + \frac{\gamma_2}{\gamma_1 + \gamma_2}(\mu_{q(z_{2N})} - \mu_{q(z_{2i})})\mu_{q(\mathbf{f}_2)}) -$$
$$\sigma^2_{q(z_{0i})}(\sum_{j=1}^{N-1} \mu_{q(z_{0j})}\mathbb{1}\{i \neq j\}\mathbf{1}_p)$$

$$\sigma^2_{q(z_{1i})} = (\mu_{q(\sigma^{-2}_{z_1})} + tr((\boldsymbol{\Sigma}_{q(\mathbf{f}_1)} + \mu_{q(\mathbf{f}_1)}\mu'_{q(\mathbf{f}_1)})(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}))^{-1}$$

$$\mu_{q(z_{1i})} = \sigma^2_{q(z_{1i})}\Big(\mu'_{q(\mathbf{f}_1)}(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i(\mathbf{h}_i) - (\mu_{q(z_{0i})}\mathbf{1}_p + \frac{\gamma_2}{\gamma_1 + \gamma_2}\mu_{q(z_{2i})}\boldsymbol{\mu}_{q(\mathbf{f}_2)}))\Big)$$

$$\sigma^2_{q(z_{2i})} = (\mu_{q(\sigma^{-2}_{z_2})} + \frac{\gamma_2^2}{\gamma_1 + \gamma_2}tr((\boldsymbol{\Sigma}_{q(\mathbf{f}_2)} + \mu_{q(\mathbf{f}_2)}\mu'_{q(\mathbf{f}_2)})\boldsymbol{\Sigma}^{-1}))^{-1}$$

$$\mu_{q(z_{2i})} = \sigma^2_{q(z_{2i})}\Big(\mu'_{q(\mathbf{f}_2)}\gamma_2\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i(\mathbf{h}_i) - (\mu_{q(z_{0i})}\mathbf{1}_p + \mu_{q(z_{1i})}\boldsymbol{\mu}_{q(\mathbf{f}_1)}))\Big)$$

$$d_{q(\eta_\mathbf{f})} = d + 1/2 * tr(\mathbf{P}_1^-(\Sigma_{q(\mathbf{f}_1)} + \mu_{q(\mathbf{f}_1)}\mu'_{q(\mathbf{f}_1)} + \Sigma_{q(\mathbf{f}_2)} + \mu_{q(\mathbf{f}_2)}\mu'_{q(\mathbf{f}_2)}))$$

$$d_{q(\lambda_\mathbf{f})} = d + 1/2 * tr(\mathbf{P}_2^-(\Sigma_{q(\mathbf{f}_1)} + \mu_{q(\mathbf{f}_1)}\mu'_{q(\mathbf{f}_1)} + \Sigma_{q(\mathbf{f}_2)} + \mu_{q(\mathbf{f}_2)}\mu'_{q(\mathbf{f}_2)}))$$

$$b_{q(\sigma^2_{z_0})} = b + 1/2\sum_{i=1}^{N-1}(\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})})$$

$$b_{q(\sigma^2_{z_1})} = b + 1/2\sum_{i=1}^{N}(\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})$$

$$b_{q(\sigma^2_{z_2})} = b + 1/2\sum_{i=1}^{N}(\sigma^2_{q(z_{2i})} + \mu^2_{q(z_{2i})})$$

C.4 Convergence Criterion

The adapted variational Bayes algorithm is run until changes in $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})]$ are below a certain threshhold. This value can be computed in each iteration as follows.

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ (f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}}) f(\boldsymbol{\theta}_{-\mathbf{w}})) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}}) + log\ f(\boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}})] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{f}_1) - log\ q(\mathbf{f}_1)] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{f}_2) - log\ q(\mathbf{f}_2)] \\
&\quad + \sum_{i=1}^{(N-1)} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(z_{0i}) - log\ q(z_{0i})] \\
&\quad + \sum_{i=1}^{N} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(z_{1i}) - log\ q(z_{1i})] \\
&\quad + \sum_{i=1}^{N} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(z_{2i}) - log\ q(z_{2i})] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\sigma^2_{z_0}) - log\ q(\sigma^2_{z_0})] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\sigma^2_{z_1}) - log\ q(\sigma^2_{z_1})] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\sigma^2_{z_2}) - log\ q(\sigma^2_{z_2})] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\eta_f) - log\ q(\eta_f)] \\
&\quad + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\lambda_f) - log\ q(\lambda_f)]
\end{aligned}
$$

Now looking at each piece individually,

$$
\begin{aligned}
& E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{X}, \mathbf{w} \mid \boldsymbol{\theta}_{-\mathbf{w}})] \\
= \quad & E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\sum_{i=1}^{N} (log[(2\pi)^{-p/2} \mid (\gamma_1 + \gamma_2)^{-1}\boldsymbol{\Sigma} \mid^{-1/2}])\Big] \\
& + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\sum_{i=1}^{N} -\frac{1}{2}[(\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{X}_i(\mathbf{h}_i) - 2\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}(z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1 + \frac{\gamma_2}{\gamma_1 + \gamma_2} z_{2i}\mathbf{f}_2) + \\
& (z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1 + \frac{\gamma_2}{\gamma_1 + \gamma_2} z_{2i}\mathbf{f}_2)'(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}(z_{0i}\mathbf{1}_p + z_{1i}\mathbf{f}_1 + \frac{\gamma_2}{\gamma_1 + \gamma_2} z_{2i}\mathbf{f}_2)]\Big] \\
= \quad & \sum_{i=1}^{N} (log[(2\pi)^{-p/2} \mid (\gamma_1 + \gamma_2)^{-1}\boldsymbol{\Sigma} \mid^{-1/2}]) \\
& + \Big[\sum_{i=1}^{N} -\frac{1}{2}\Big(\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{X}_i(\mathbf{h}_i) - \\
& 2\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}\mu_{q(z_{0i})}\mathbf{1}_p - 2\mathbf{X}_i(\mathbf{h}_i)'(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}\mu_{q(z_{1i})}\boldsymbol{\mu}_{q(\mathbf{f}_1)} - \\
& 2\mathbf{X}_i(\mathbf{h}_i)'\gamma_2\boldsymbol{\Sigma}^{-1}\mu_{q(z_{2i})}\boldsymbol{\mu}_{q(\mathbf{f}_2)} + \\
& (\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})tr((\boldsymbol{\Sigma}_{q(\mathbf{f}_1)} + \boldsymbol{\mu}_{q(\mathbf{f}_1)}\boldsymbol{\mu}'_{q(\mathbf{f}_1)})(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}) + \\
& (\sigma^2_{q(z_{2i})} + \mu^2_{q(z_{2i})})tr((\boldsymbol{\Sigma}_{q(\mathbf{f}_2)} + \boldsymbol{\mu}_{q(\mathbf{f}_2)}\boldsymbol{\mu}'_{q(\mathbf{f}_2)})\frac{\gamma_2^2}{(\gamma_1 + \gamma_2)}\boldsymbol{\Sigma}^{-1}) + \\
& 2\mu_{q(z_{0i})}\mu_{q(z_{1i})}\mathbf{1}'_p(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_1)} + 2\mu_{q(z_{1i})}\mu_{q(z_{2i})}\boldsymbol{\mu}'_{q(\mathbf{f}_1)}\gamma_2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_2)} + \\
& 2\mu_{q(z_{0i})}\mu_{q(z_{2i})}\mathbf{1}'_p\gamma_2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{q(\mathbf{f}_2)}\Big)\Big] - \\
& \Big[\sum_{i=1}^{N-1}(\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})}) + \frac{1}{2}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\mu_{q(z_{0i})}\mu_{q(z_{0j})}\mathbb{1}\{j \neq i\}\Big]\mathbf{1}'_p(\gamma_1 + \gamma_2)\boldsymbol{\Sigma}^{-1}\mathbf{1}_p
\end{aligned}
$$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{f}_1) - log\ q(\mathbf{f}_1)] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[- \frac{p}{2} log\ 2\pi + \frac{1}{2} log\ \mid \eta_f \mathbf{P}_1^- + \lambda_f \mathbf{P}_2^- \mid \Big] - \\
& \quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}(tr[\mathbf{f}_1 \mathbf{f}_1' (\eta_f \mathbf{P}_1^- + \lambda_f \mathbf{P}_2^-)]\Big] + \\
& \quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{p}{2} log\ 2\pi + \frac{1}{2} log\ \mid \boldsymbol{\Sigma}_{q(\mathbf{f}_1)} \mid \Big] + \\
& \quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2} tr(\mathbf{f}_1 \mathbf{f}_1' \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}^{-1}) - \mathbf{f}_1' \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}^{-1} \boldsymbol{\mu}_{q(\mathbf{f}_1)}\Big] + \\
& \quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2} \boldsymbol{\mu}_{q(\mathbf{f}_1)}' \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}^{-1} \boldsymbol{\mu}_{q(\mathbf{f}_1)}\Big] \\
&= C + \frac{1}{2} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[2log\ \eta_f] + \frac{1}{2} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[(p-2) log\ \lambda_f] - \\
& \quad \frac{1}{2} tr\Big((\boldsymbol{\Sigma}_{q(\mathbf{f}_1)} + \boldsymbol{\mu}_{q(\mathbf{f}_1)} \boldsymbol{\mu}_{q(\mathbf{f}_1)}')(\mu_{q(\eta_f)} \mathbf{P}_1^- + \mu_{q(\lambda_f)} \mathbf{P}_2^-)\Big) - \\
& \quad \frac{1}{2} log\ \mid \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}^{-1} \mid + \frac{p}{2}
\end{aligned}
$$

where C is a constant that does not change from one iteration to the next. Similarly,

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{f}_2) - log\ q(\mathbf{f}_2)] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[- \frac{p}{2} log\ 2\pi + \frac{1}{2} log\ \mid \eta_f \mathbf{P}_1^- + \lambda_f \mathbf{P}_2^- \mid \Big] - \\
& \quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2}(tr[\mathbf{f}_2 \mathbf{f}_2' (\eta_f \mathbf{P}_1^- + \lambda_f \mathbf{P}_2^-)]\Big] + \\
& \quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{p}{2} log\ 2\pi + \frac{1}{2} log\ \mid \boldsymbol{\Sigma}_{q(\mathbf{f}_2)} \mid \Big] + \\
& \quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2} tr(\mathbf{f}_2 \mathbf{f}_2' \boldsymbol{\Sigma}_{q(\mathbf{f}_2)}^{-1}) - \mathbf{f}_2' \boldsymbol{\Sigma}_{q(\mathbf{f}_2)}^{-1} \boldsymbol{\mu}_{q(\mathbf{f}_2)}\Big] + \\
& \quad E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{1}{2} \boldsymbol{\mu}_{q(\mathbf{f}_2)}' \boldsymbol{\Sigma}_{q(\mathbf{f}_2)}^{-1} \boldsymbol{\mu}_{q(\mathbf{f}_2)}\Big] \\
&= C + \frac{1}{2} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[2log\ \eta_f] + \frac{1}{2} E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[(p-2) log\ \lambda_f] - \\
& \quad \frac{1}{2} tr\Big((\boldsymbol{\Sigma}_{q(\mathbf{f}_2)} + \boldsymbol{\mu}_{q(\mathbf{f}_2)} \boldsymbol{\mu}_{q(\mathbf{f}_1)}')(\mu_{q(\eta_f)} \mathbf{P}_1^- + \mu_{q(\lambda_f)} \mathbf{P}_2^-)\Big) - \\
& \quad \frac{1}{2} log\ \mid \boldsymbol{\Sigma}_{q(\mathbf{f}_1)}^{-1} \mid + \frac{p}{2}
\end{aligned}
$$

where C is a constant that does not change from one iteration to the next. For $\mathbf{z}_0$ $= (z_{01}, \ldots, z_{0(N-1)})'$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{z}_0) - log\ q(\mathbf{z}_0)] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{N-1}{2} log\ 2\pi - \frac{N-1}{2} log\ \sigma^2_{z_0} - \sum_{i=1}^{N-1} -\frac{1}{2\sigma^2_{z_0}} z^2_{0i} + \\
&\quad \frac{N-1}{2} log\ 2\pi + \frac{N-1}{2} log\ \sigma^2_{q(z_{0i})} + \\
&\quad \sum_{i=1}^{N-1} \frac{1}{2\sigma^2_{q(z_{0i})}} (z_{0i} - \mu_{q(z_{0i})})^2 \Big] \qquad \text{(C.2)} \\
&= \frac{N-1}{2} log\ \sigma^2_{q(z_{0i})} - E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{N-1}{2} log\ \sigma^2_{z_0}\Big] - \qquad \text{(C.3)} \\
&\quad \frac{1}{2}\mu_{q(\frac{1}{\sigma^2_{z_0}})}\Big(\sum_{i=1}^{N-1} (\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})})\Big) + \frac{N-1}{2}
\end{aligned}
$$

For $\mathbf{z}_1 = (z_{11}, \ldots, z_{1N})'$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{z}_1) - log\ q(\mathbf{z}_1)] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{N}{2} log\ 2\pi - \frac{N}{2} log\ \sigma^2_{z_1} - \sum_{i=1}^{N} -\frac{1}{2\sigma^2_{z_1}} z^2_{1i} + \\
&\quad \frac{N}{2} log\ 2\pi + \frac{N}{2} log\ \sigma^2_{q(z_{1i})} + \sum_{i=1}^{N} \frac{1}{2\sigma^2_{q(z_{1i})}} (z_{1i} - \mu_{q(z_{1i})})^2 \Big] \\
&= \frac{N}{2} log\ \sigma^2_{q(z_{1i})} - E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{N}{2} log\ \sigma^2_{z_1}\Big] - \\
&\quad \frac{1}{2}\mu_{q(\frac{1}{\sigma^2_{z_1}})}\Big(\sum_{i=1}^{N} (\sigma^2_{q(z_{1i})} + \mu^2_{q(z_{1i})})\Big) + \frac{N}{2}
\end{aligned}
$$

For $\mathbf{z}_2 = (z_{21}, \ldots, z_{2N})'$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{z}_2) - log\ q(\mathbf{z}_2)] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[-\frac{N}{2} log\ 2\pi - \frac{N}{2} log\ \sigma^2_{z_2} - \sum_{i=1}^{N} -\frac{1}{2\sigma^2_{z_2}} z^2_{2i} + \\
&\quad \frac{N}{2} log\ 2\pi + \frac{N}{2} log\ \sigma^2_{q(z_{2i})} + \sum_{i=1}^{N} \frac{1}{2\sigma^2_{q(z_{2i})}} (z_{2i} - \mu_{q(z_{2i})})^2 \Big] \\
&= \frac{N}{2} log\ \sigma^2_{q(z_{2i})} - E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[\frac{N}{2} log\ \sigma^2_{z_2}\Big] - \\
&\quad \frac{1}{2}\mu_{q(\frac{1}{\sigma^2_{z_2}})}\Big(\sum_{i=1}^{N} (\sigma^2_{q(z_{2i})} + \mu^2_{q(z_{2i})})\Big) + \frac{N}{2}
\end{aligned}
$$

For $\sigma^2_{z_0}$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\sigma^2_{z_0}) - log\ q(\sigma^2_{z_0})] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{b^a}{\Gamma(a)} - (a+1)log\ \sigma^2_{z_0} - b\frac{1}{\sigma^2_{z_0}} - \\
&\quad log\ \frac{b^{a_{q(\sigma^2_{z_0})}}_{q(\sigma^2_{z_0})}}{\Gamma(a_{q(\sigma^2_{z_0})})} + (a_{q(\sigma^2_{z_0})}+1)log\ \sigma^2_{z_0} + \\
&\quad b_{q(\sigma^2_{z_0})}\frac{1}{\sigma^2_{z_0}}\Big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[-(a+1)log\ \sigma^2_{z_0}] - b\mu_{q(\frac{1}{\sigma^2_{z_0}})} - log\ \frac{b^{a_{q(\sigma^2_{z_0})}}_{q(\sigma^2_{z_0})}}{\Gamma(a_{q(\sigma^2_{z_0})})} + \\
&\quad log\ \frac{b^a}{\Gamma(a)} + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[(a_{q(\sigma^2_{z_0})}+1)log\ \sigma^2_{z_0}] + b_{q(\sigma^2_{z_0})}\mu_{q(\frac{1}{\sigma^2_{z_0}})}
\end{aligned}
$$

For $\sigma^2_{z_1}$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\sigma^2_{z_1}) - log\ q(\sigma^2_{z_1})] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{b^a}{\Gamma(a)} - (a+1)log\ \sigma^2_{z_1} - b\frac{1}{\sigma^2_{z_1}} - \\
&\quad log\ \frac{b^{a_{q(\sigma^2_{z_1})}}_{q(\sigma^2_{z_1})}}{\Gamma(a_{q(\sigma^2_{z_1})})} + (a_{q(\sigma^2_{z_1})}+1)log\ \sigma^2_{z_1} + \\
&\quad b_{q(\sigma^2_{z_1})}\frac{1}{\sigma^2_{z_1}}\Big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[-(a+1)log\ \sigma^2_{z_1}] - b\mu_{q(\frac{1}{\sigma^2_{z_1}})} - log\ \frac{b^{a_{q(\sigma^2_{z_1})}}_{q(\sigma^2_{z_1})}}{\Gamma(a_{q(\sigma^2_{z_1})})} + \\
&\quad log\ \frac{b^a}{\Gamma(a)} + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[(a_{q(\sigma^2_{z_1})}+1)log\ \sigma^2_{z_1}] + b_{q(\sigma^2_{z_1})}\mu_{q(\frac{1}{\sigma^2_{z_1}})}
\end{aligned}
$$

For $\sigma^2_{z_2}$

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\sigma^2_{z_2}) - log\ q(\sigma^2_{z_2})] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{b^a}{\Gamma(a)} - (a+1)log\ \sigma^2_{z_2} - b\frac{1}{\sigma^2_{z_2}} - \\
&\quad log\ \frac{b^{a_{q(\sigma^2_{z_2})}}_{q(\sigma^2_{z_2})}}{\Gamma(a_{q(\sigma^2_{z_2})})} + (a_{q(\sigma^2_{z_2})}+1)log\ \sigma^2_{z_2} + \\
&\quad b_{q(\sigma^2_{z_2})}\frac{1}{\sigma^2_{z_2}}\Big] \\
&= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[-(a+1)log\ \sigma^2_{z_2}] - b\mu_{q(\frac{1}{\sigma^2_{z_2}})} - log\ \frac{b^{a_{q(\sigma^2_{z_2})}}_{q(\sigma^2_{z_2})}}{\Gamma(a_{q(\sigma^2_{z_2})})} + \\
&\quad log\ \frac{b^a}{\Gamma(a)} + E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[(a_{q(\sigma^2_{z_2})}+1)log\ \sigma^2_{z_2}] + b_{q(\sigma^2_{z_2})}\mu_{q(\frac{1}{\sigma^2_{z_2}})}
\end{aligned}
$$

For η_f

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\eta_f) - log\ q(\eta_f)] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{d^c}{\Gamma(c)} + (c-1)log\ \eta_f - d\eta_f - \\
&\quad log\ \frac{d^{c_{q(\eta_f)}}_{q(\eta_f)}}{\Gamma(c_{q(\eta_f)})} - c\ log\ \eta_f + d_{q(\eta_f)}\eta_f\Big] \\
&= log\ \frac{d^c}{\Gamma(c)} - log\ \frac{d^{c_{q(\eta_\chi)}}_{q(\eta_f)}}{\Gamma(c_{q(\eta_f)})} - 2E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ \eta_f] - d\mu_{q(\eta_f)} + \\
&\quad d_{q(\eta_f)}\mu_{q(\eta_f)}
\end{aligned}
$$

For λ_f

$$
\begin{aligned}
E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\lambda_f) - log\ q(\lambda_f)] &= E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}\Big[log\ \frac{d^c}{\Gamma(c)} + (c-1)log\ \lambda_f - d\lambda_f - \\
&\quad log\ \frac{d^{c_{q(\lambda_f)}}_{q(\lambda_f)}}{\Gamma(c_{q(\lambda_f)})} - \Big(\frac{p-2}{2} + c - 1\Big)\ log\ \lambda_f + d_{q(\lambda_f)}\lambda_f\Big] \\
&= log\ \frac{d^c}{\Gamma(c)} - log\ \frac{d^{c_{q(\lambda_f)}}_{q(\lambda_f)}}{\Gamma(c_{q(\lambda_f)})} - (p-2)E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ \lambda_f] - d\mu_{q(\lambda_f)} + \\
&\quad d_{q(\lambda_f)}\mu_{q(\lambda_f)}
\end{aligned}
$$

The expression for $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{X}, \mathbf{w}, \boldsymbol{\theta}_{-\mathbf{w}}) - log\ q(\boldsymbol{\theta}_{-\mathbf{w}})]$ can be simplified much further by combining terms that cancel out. However, in some cases the ability to cancel terms depends on the order of the updates. For instance, in the expression, $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\sigma^2_{z_0}) - log\ q(\sigma^2_{z_0})]$, the terms $-b\mu_{q(\frac{1}{\sigma^2_{z_0}})}$ and $b_{q(\sigma^2_{z_0})}\mu_{q(\frac{1}{\sigma^2_{z_0}})}$ cancel with $-\frac{1}{2}\mu_{q(\frac{1}{\sigma^2_{z_0}})}\Big(\sum_{i=1}^{N-1}(\sigma^2_{q(z_{0i})} + \mu^2_{q(z_{0i})})\Big)$ from $E_{q(\boldsymbol{\theta}_{-\mathbf{w}})}[log\ f(\mathbf{z}_0) - log\ q(\mathbf{z}_0)]$ as long as the parameters of $q(\mathbf{z}_0)$ are updated before $b_{q(\sigma^2_{z_0})}$. For convenience, we have taken account the ordering necessary to compute the convergence criterion in the updates given above. Additionally, note all components in this expression that do not change from one iteration to the next can be ignored.

BIBLIOGRAPHY

[1] Behseta,S., Kass, R., and Wallstrom, G.(2005). Hierarchical models for assessing variability among functions. *Biometrika* **92**, 2, 419-434.

[2] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

[3] Boudaoud, S., Rix, H., Meste, O. (2010). Core shape modelling of a set of curves. *Computational Statistics and Data Analysis* **54**, 2, 308-325.

[4] Brumback, C.L., and Lindstrom, J.M. (2004). Self-modeling with flexible, random time transformations. *Biometrics* **60**, 461-470.

[5] Cai, T., and Yuan, M. (2010). Nonparametric covariance function estimation for functional and longitudinal data. *Technical Report*.

[6] Calderhead, B., Girolami, M., and Lawrence, N.(2009). Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. *Advances in neural information processing systems* **21**, 217-224.

[7] Carey, J., Liedo, P., Müller, H. G., Wang, J.L., and Chiou, J.M. (1998). Relationship of age patterns of fecundity to mortality, longevity and lifetime reproduction in a large cohort of Mediterranean fruit fly females. *J. Gerontol* **A 53**, B245-B251.

[8] Dawid, A.P. (1981). Matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* **68**, 1, 265-274.

[9] Earls, C., and Hooker, G. (2014). Bayesian covariance estimation and inference in latent Gaussian process models. *Statistical Methodology* **18**, 79-100.

[10] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models(comment on article by Browne and Draper). *Bayesian Analysis* **1**, 3, 515-534.

[11] Crainiceanu, C., and Goldsmith, J. (2010). Bayesian Functional Data Analysis Using WinBUGS. *Journal of Statistical Software* **32**, 1-33, 835-837.

[12] Ferraty, F., and Vieu, P. (2006). *Nonparameteric Functional Data Analysis: Theory and Practice.* Springer, New York.

[13] Gervini, D., and Gasser, T. (2004). Self-modeling warping functions. *Journal of the Royal Statistical Society, Ser. B* **66**, 959-971.

[14] Gervini, D., and Gasser, T. (2005). Nonparametric maximum likelihood estimation of the structure of a sample of curves. *Biometrika* **92**, 801-820.

[15] Goldsmith, J. , Bobb, J., Crainiceanu, C., Caffo, B., and Reich, D. (2010). Penalized functional regression. *Journal of Computational and Graphical Statistics* **20**, 4, 830-851.

[16] Goldsmith, J., Wand, M.P., and Crainiceanu, C.(2011). Functional regression via variational Bayes. *Electronic Journal of Statistics* **5**, 572.

[17] James, G.M.(2007). Curve alignment by moments. *The Annals of Applied Statistics* **1**, 2, 480-501.

[18] Kauermann, G., and Wegener, M. (2009). Functional variance estimation using penalized splines with principal component analysis. *Statistics and Computing* **21**, 2, 159-171.

[19] Kaufman, C., and Sain, S. (2010). Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis* **5**, 1, 123-150.

[20] Kneip, A., and Gasser, T. (1992) Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics* **1**, 2, 480-501.

[21] Kneip, A., and Gasser, T. (1995) Searching for structure in curve samples. *Journal of the American Statistical Association* **90**, 1179-1188.

[22] Kneip, A., and Ramsay J.O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association* **103**, 483, 1155-1165.

[23] Kullback, S., and Leibler, D.(1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79-86.

[24] Latuszyński, K., Roberts, G., and Rosenthal, J.(2011). Adaptive Gibbs samplers and related MCMC methods. *Arxiv preprint* **arXiv:1101.5838**.

[25] Linde, A. (2011). Reduced rank regression models with latent variables in Bayesian functional data analysis. *Bayesian Analysis* **6**, 1, 77-126.

[26] Liu, X., andMüller, H.G.(2004). Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association* **99**, 687-699.

[27] Liu, X., and Yang, M.C.K. (2009). Simultaneous curve registration and clustering for functional data. *Computational Statistics and Data Analysis* **53**, 1361-1376.

[28] Müller, H.G., and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics* **33**, 2, 774-805.

[29] Nguyen, X., and Gelfand A. E. (2011). The dirichlet labeling process for clustering functional data. *Statistica Sinica* **21**, 1249-1289.

[30] Omerod, J., and Wand, M. (2010). Explaining variational approximations. *The American Statistician* **64**, 140-153.

[31] Rakêt, L, Sommer, S., and Markussen, B. (2014). A nonlinear mixed-effects model for simultaneous smoothing and registration of functional data. *Pattern Recognition Letters* **38**, 1-7.

[32] Ramsay, J., and Silverman, B. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* Springer, New York.

[33] Ramsay, J.O., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and Matlab.* Springer Science, New York.

[34] Ramsay, J.O., and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society.Series B (Statistical Methodology)* **60**, 2, 351-363.

[35] Ramsay, J., and Silverman, B. (2005). *Applied Functional Data Analysis.* Springer-Verlag, New York.

[36] Ramsay, J., and Silverman, B. (2005). *Functional Data Analysis.* Springer, New York.

[37] Rao, C.R.(1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society* **37**, 81-91.

[38] Ronn, B. (2001) Nonparameteric maximum likelihood estimation of shifted curves. *Journal of the Royal Statistical Society, B* **63**, 243-259.

[39] Sakoe, H., and Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* **26**,1, 43-49.

[40] Sangalli, L.M., Secchi, P., Vantini, S., and Vitelli, V.(2010). k-mean alignment for curve clustering. *Computational Statistics and Data Analysis*.**54**, 1219-1233.

[41] Silverman, B.W.(1995). Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society*.**57**, 673-689.

[42] Silverstein, J. (1985). The smallest eigenvalue of a large dimensional Wishart matrix. *The Annals of Probability* **13**, 4, 1364-1368.

[43] Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J.S. (2011). Registration of functional data using fisher-rao metric. *arXiv preprint arXiv:1103.3817*

[44] Tang, R., and Müller, H.G.(2008). Pairwise curve synchronization for functional data. *Biometrika* **95**, 4, 875-889.

[45] Telesca, D., and Inoue, L.Y.T. (2007). Bayesian hierarchical curve registration. *Journal of the American Statistical Association* **103**, 481, 328-339.

[46] Tuddenham, R.D., and Snyder, M.M. (1954). Physical growth of California boys and girls from birth to eighteen years. *University of California Publications in Child Development I*, 183-364.

[47] Wahba, G. (1992). *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics.

[48] Wang, K., and Gasser, T. (1997). Alignment of curves by dynamic time warping. *The Annals of Statistics* **25**, 3, 1251-1276.

[49] El-niño. *Wikipedia, The Free Encyclopedia* Wikipedia, The Free Encyclopedia. 6 May. 2014. Web. 8 May. 2014.

[50] Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society* **73**, 1, 3-36.

[51] Yao, F., Müller, H.G., Wang, J.L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100(470)**, 577-590.

[52] Zhang, Y., and Telesca, D.(2014). Joint clustering and registration of functional data. *arXiv preprint arXiv:1403.7134*.